Raising the bar for reading comprehension

van Kuijk, Mechteld Femke

Raising the bar for reading comprehension

The effects of a teacher professional development program targeting goals, data use, and instruction

Mechteld van Kuijk

ico
Interuniversity Center for Educational Research

ISBN: 978-90-367-6774-3 (printed version)

ISBN: 978-90-367-6852-8 (electronic version)

Cover design: Evert Wilstra

Foto: Hilbert Geerling

Printed by: Gildeprint Drukkerijen

rijksuniversiteit
groningen

Raising the bar for reading comprehension

The effects of a teacher professional development program
targeting goals, data use, and instruction

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 13 maart 2014 om 16.15 uur

door

Mechteld Femke van Kuijk
geboren op 29 augustus 1985
te Kingston upon Thames, Verenigd Koninkrijk

Promotor

Prof. dr. R.J. Bosker

Copromotor

Dr. M.I. Deunk

Beoordelingscommissie

Prof. dr. D. Muijs

Prof. dr. W.J.C.M. van de Grift

Prof. dr. F.J.G. Janssens

Table of contents

Chapter 1: General introduction 7

Chapter 2: The effect of the professional development program on reading comprehension 17

Chapter 3: Investigating the validity of cutscores 37

Chapter 4: Teacher-set performance goals and relations to student achievement 61

Chapter 5: Exploring teacher implementation of the professional development program 85

Chapter 6: General conclusion and discussion 109

Appendices 119

Samenvatting (Dutch summary) 171

References 181

About the author 197

Dankwoord (acknowledgements in Dutch) 199

ICO Dissertation series 203

1. General introduction

1.1 Research background

Reading comprehension is an important basis for further learning, working, and living as a large amount of information is transferred via printed or digital text (Kirsch, 2002; Reis, McCoach, Little, Muller, & Kaniskan, 2011; Snow, Burns, & Griffin, 1998; Van Elsäcker, 2002). Yet reading comprehension is a skill which is difficult to master as comprehension stems from an active and interactive process between the *reader* (with a specific level of e.g., decoding skills, vocabulary and motivation), the specific *text* (with certain characteristics with respect to e.g., text genre, audience appropriateness and coherence), and the *goal* a reader has for that specific text (Snow, 2002; Sweet & Snow, 2003). The rationale for explicit teaching of reading comprehension in primary schools is that comprehension can be improved by teaching students to use specific strategies or to reason strategically when they encounter difficulties in understanding what they are reading (National Reading Panel, 2000; Pressley, 1998). The long-term effects of early acquired proficiency in reading have been widely established in the literature: being a proficient reader at an early age is a predictor of later academic success (e.g. Bodovski & Youn, 2011; Snow et al., 1998) and is related to lower levels of grade retention (Jimerson & Kaufman, 2003), high school drop-out (Lloyd, 1978), and delinquent behavior (Stattin & Klackenberg-Larsson, 1993).

Currently, there are performance concerns in the Netherlands pertaining to the reading results of students in primary schools (e.g., Ministry of Education, 2008; 2010). For example, on the Dutch periodical assessment of educational achievement (known as PPON), 30 percent of the third-grade students read at a level which, according to reading experts and teachers, should be attainable for 75 percent (Van Berkel et al., 2007). Furthermore, although the scores on the 2011 international PIRLS assessment (targeting fourth-grade reading) indicate that, comparatively speaking, students in the Netherlands perform rather well, the average achievement of the Dutch students is significantly lower than in 2001 (Meelissen et al., 2012). The national performance concerns pertain particularly to the degree to which struggling, poorly performing readers are prepared for later schooling and the work force (Inspectorate of Education, 2007; 2010b). On the 2012 international PISA test (targeting - among other areas - the reading skills of 15-year olds), it is found that almost 14 percent of Dutch students demonstrate such low levels of literacy that they are considered to have difficulties participating in society (Kordes, Bolsinova, Limpens, & Stolwijk, 2013). As "[r]eading is essential to our success in society" (Snow et al., 1998, p. 17), the improvement of student reading comprehension is a priority for Dutch policymakers and practitioners; therefore its place on the research agenda is obvious.

Characteristics of effective teacher instruction that foster the development of students' reading skills have been identified in various studies (discussed in, e.g., National Reading Panel,

2000). Yet in the Netherlands, as in many other countries, there is a gap between the findings of empirical research and the actual instruction provided by teachers (Aarnoutse & Weterings, 1995; Andreassen & Braten, 2011; Liang & Dole, 2006; Van Keer & Verhaeghe, 2005). Researchers in the area of reading have emphasized that schools and teachers should operate on a "what works" basis, in which scientific evidence is used to help improve teacher practice and student performance (e.g., Armbruster, Lehr, & Osborn, 2010; Collins Block & Lacina, 2009; National Reading Panel, 2000; Pressley, 1998).

In order to implement new (i.e., more effective) instructional techniques and to change existing routines, teachers need support and guidance (Black & Wiliam, 1998b; Borko, 2004). For this purpose, teacher Professional Development (PD) programs are frequently used. Teacher PD programs are "systematic efforts to bring about change in the classroom practices of teachers, in their attitudes and beliefs, and in the learning outcomes of students" (Guskey, 2002, p. 381). Recent educational reforms, implemented around the world to foster student learning, rely heavily on teacher learning and improved instruction to increase student performance (Borko, Borko, & Koellner, 2010; Desimone, 2009; Garet, Porter, Desimone, Birman, & Yoon, 2001; Guskey, 2002; Hill, 2007; OECD, 2005). As support has been gained for the view that Dutch students' reading comprehension can be improved by targeting the teachers' instructional practices, we developed a teacher PD program. This program was designed following the tradition of *applied research* – a research paradigm which aims to produce knowledge for the solution of a practical problem (McMillan & Schumacher, 1989).

1.2 Content of the professional development program

The teacher PD program was designed aiming to improve Dutch students' reading comprehension via the training and coaching of second and third grade teachers. These grades, with students of approximately 7 to 9 years old, were specifically targeted due to the importance of early acquired reading proficiency and the fact that the most promising results of interventions and reforms are found in the junior grades (Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988). In the Netherlands, reading comprehension is a separate subject in the curriculum from the second grade onward, and studies have shown that reading comprehension can successfully be taught in these early grades of primary school (e.g., Aarnoutse, 1991; Van Elsäcker, 2002).

The rationale behind the PD program was that students' reading comprehension was expected to improve by making teachers' instruction more goal-oriented, focused, clear, and better suited to students' needs. The teachers that participated in the PD program were supported in improving their practice with help of a three-component program: 1) setting standards and performance goals for every student, 2) applying formative assessment and data use, and 3) acquiring relevant instructional skills and (content and curriculum) knowledge in reading comprehension. All three

components have shown to be positively related to student performance; these components are discussed in the following paragraphs.

1.2.1. Component 1: Setting standards and performance goals for every student

Goal setting was incorporated as the first component in the PD program as the insufficient results of Dutch students on both international and national assessments had been attributed to the fact that, for schools and teachers, it was unclear what students should know and do at certain time points (Expert group Continuous Learning Progression, 2008). Clearly defined performance goals were desired according to several educational authorities in the Netherlands (Council of Education, 2007; Inspectorate of Education, 2011; Ministry of Education, 2010) as these goals were assumed to make teacher instruction more targeted which, subsequently, was assumed to result in improved student outcomes. Working with goals has generally been proven to be effective for enhancing performance. Setting goals leads to a clearer notion of how success can be attained and it focuses the attention on the realization of relevant outcomes (e.g., Fuchs, Fuchs, & Deno, 1985; Locke & Latham, 1990). Particularly goals that are set at an ambitious level are associated with higher outcomes (Locke & Latham, 1990; 2002).

As part of the PD program, we asked the teachers to set a performance goal for each of their students. The goals were formulated by selecting one of five *performance categories* (labeled below minimum, minimum, basic, proficient, and advanced), in acknowledgement of the differences between students' capabilities. The performance categories were defined by the participating teachers in an early stage of the PD program with help of a *standard setting procedure*. In this procedure, the participants considered items from reading comprehension assessments and discussed their performance expectations for students with different cognitive abilities. After several rounds of standard setting, the performance categories were established. Following this procedure, the participating teachers formulated performance goals for each individual student in their class by selecting one of the five performance categories: e.g., 'At the end of the school year, I want Billy to perform at the *proficient* category and Jenny to perform at the *advanced* category'. The student-specific goals were frequently re-examined and referred to during the course of the PD program. At the end of the school year, the participating teachers received an overview of the degree to which the performance goals had been attained.

As it was important that these goals were set at an appropriate level for each student given their capabilities, we developed a *multistep procedure* which incorporated performance data analysis and team discussion to help teachers reflect on and reconsider the goals' appropriateness before deciding on its final version - following recommendations of the data use literature (e.g., Schildkamp & Kuiper, 2010). These aspects of the multistep procedure pertain to the second component of our program.

1.2.2. Component 2: Applying formative assessment and data use

In order to help the participating teachers set appropriate goals and to help teachers attain these goals, it was important that they based their instructional decisions on assessment results (e.g., Guskey, 2002). Using student performance data to adapt one's teaching in order to better meet students' needs is known as *formative assessment* (Black & Wiliam, 1998a; 1998b; Herman, Osmundson, & Silver, 2010). In their meta-analysis, Black and Wiliam (1998b) conclude that there is "a body of firm evidence that formative assessment is an essential component of classroom work and that its development can raise (…) achievement" (p.148). Comparable results have been reported in school effectiveness research, in which a frequent evaluation and monitoring of performance is found to be positively associated with pupil achievement (Muijs & Reynolds, 2011; Sammons, Hillman, & Mortimore, 1997; Scheerens & Bosker, 1997). Other studies have shown that schools and districts applying a "data-driven" way of teaching can result in increased student performance levels (e.g., Carlson, Borman, & Robinson, 2011; Slavin et al., 2013). In the Netherlands, the progress of students is monitored with the use of *student monitoring systems*. Yet teachers use these systems to a rather limited extent for the purpose of analyzing problems and, subsequently, adapting instruction. Furthermore, teachers who do use the student monitoring systems for these purposes are often unaware of the possibilities for more sophisticated analyses (Ledoux, Blok, & Boogaard, 2009; Meijer & Ledoux, 2011; Schildkamp & Kuiper, 2010; van der Kleij & Eggen, 2013). The limited use of performance data and the student monitoring systems is particularly relevant given the finding that teachers hardly differentiate. *Differentiation* is defined as "an approach to teaching in which teachers proactively modify curricula, teaching methods, resources, learning activities and student products to address the diverse needs of individual students and small groups of students to maximize the learning opportunity for each student in a classroom" (Tomlinson et al., 2003, p. 121). Having better insight in the knowledge and skills that students already have attained and identifying which knowledge and skills they still need to master will benefit the degree to which instruction suits the needs of these students. Currently, only 50 percent of teachers in primary education sufficiently target different students' needs (Inspectorate of Education, 2012). The limited implementation of differentiation is particularly evident during reading comprehension lessons (Van Berkel et al., 2007).

During the PD program, the participating teachers received training in the use of the student monitoring system and interpretation of its results. We used a cyclical theoretical model to illustrate how to work with student performance data, similar to the study of Schnellert, Butler and Higginson (2008); teachers were stimulated to work in reflective cycles of goal setting, planning, teaching and monitoring. These aspects are important elements in teachers' learning process and the realization of change (Borko et al., 2010). By working with student-specific performance goals which have been set at different performance levels, and by monitoring

performance in relation to these goals - i.e., the first and second component of our program-, it was expected that differentiation would be fostered as a result. An important prerequisite, however, is that teachers adjust their practices accordingly after analyzing the data, a step which is not always guaranteed (Goertz, Olah & Riggan, 2009 in Carlson et al., 2011). Analyzing data "(…) is not enough to produce gains in achievement. Schools must actually take action to change teaching and learning" (Slavin et al., 2013, p. 390). The third component of our PD program focused on how to take action after analyzing the data.

1.2.3. Component 3: Instructional skills and knowledge in reading comprehension

After setting the performance goals and identifying the progress made toward them, it was important to help the teachers attain their own objectives by ensuring that they were sufficiently knowledgeable and skilled with respect to relevant instructional practices and knowledge in reading comprehension. In order to advance students' reading proficiency, it is important that instruction in this domain is focused and clear (e.g., Andreassen & Braten, 2011; Verhoeven, 1991), but in practice, instruction does not always meet these qualifications. Lessons in reading comprehension often take the following sequence in the Netherlands. First, the students read a text either out loud or in silence. Second, a few questions about the text are discussed with the whole class, after which the students have to answer the remaining questions independently. Last, the correct answers are discussed with the whole class (Aarnoutse, 1992). Little explicit instruction is given during lessons in reading comprehension (Aarnoutse & Weterings, 1995; Van Elsäcker, 2002) and – as abovementioned – teachers hardly differentiate between students (Van Berkel et al., 2007; Van Elsäcker, 2002). It has been hypothesized, although not empirically researched, that primary school teachers in the Netherlands find reading comprehension a difficult subject to teach due to the complexity of the reading comprehension skills and the inadequacy of the curricular textbooks used in the Netherlands (Droop, van Elsäcker, & Voeten, 2012; Houtveen & Van de Grift, 2012; Stoeldraijer & Forrer, 2012). The textbooks have been criticized as being "more bulky than necessary, containing a substantial amount of material that has little or nothing to do with learning to read" (Houtveen & Van de Grift, 2012, p. 88). They also contain a large number of reading strategies, but not all of these strategies which are presented as "effective" can be supported by empirical evidence (Droop et al., 2012; Stoeldraijer & Forrer, 2012). The inadequacy of the curriculum is considered to be problematic as teachers in the Netherlands are known to follow the curricular textbooks to a very large extent (Meelissen et al., 2012).

The PD program aimed to equip the teachers with the most relevant instructional skills and knowledge in reading comprehension. Two effective instructional practices were discussed, namely *Direct Instruction*, being a teacher-centered model for instruction focused on the content and structure of a lesson (Borman, Hewes, Overman, & Brown, 2003; Muijs & Reynolds, 2011)

and *modeling*, which is an instructional technique in which the teacher demonstrates how to apply a reading strategy or solve a problem by thinking aloud and linking the solution to skills or knowledge that the students already possess (Fisher, Frey, & Lapp, 2008; National Reading Panel, 2000). Modeling in combination with Direct Instruction has been identified as an effective procedure to help struggling learners and to remediate learning disabilities, as found in the meta-analyses of Swanson and Hoskyn (1998). In addition, important determinants of reading performance and key concepts in the second- and third-grade reading comprehension curriculum were discussed. Furthermore, we focused on the curricular textbooks which were used in the participating schools: among other things, the reading strategies in these textbooks were compared to those mentioned in the guidelines of the Expertise Centre for the Dutch Language (2010).

We combined the program's three components - 1) setting standards and performance goals for every student, 2) applying formative assessment and data use, and 3) acquiring relevant instructional skills and (content and curriculum) knowledge in reading comprehension - into one synergetic package, as the components were assumed to foster the desired change in instruction in an inter-related manner.

1.3 General information on the set-up of the program and characteristics of effective teacher professional development

In this paragraph, practical information on the general set-up of the program is provided. Characteristics of PD programs that are effective in improving teacher behavior and subsequently, student results, were incorporated in our multicomponent PD program's design. Here, these characteristics, as identified in the literature on effective professional development (Desimone, 2009; Garet et al., 2001; Wayne, Yoon, Zhu, Cronen, & Garet, 2008; Yoon, Duncan, Lee, Scarloss, & Skapley, 2007) are discussed as we address several features and details of our program.

The multicomponent PD program targeted second and third grade teachers from the same school as well as the school's principal and internal support coordinator. In the school year of 2011-2012, nineteen schools in the northern part of the Netherlands participated. In total, 33 second- and third-grade teachers (teaching 451 students) participated, and the school principals and internal support coordinators of these nineteen schools took part as well. Involvement in the program entailed attending nine after-school meetings (duration: 1.5 to 2.5 hours per meeting) as well as completion of accompanying homework assignments. Participants' total time investment was scheduled for approximately 40 hours. Participation was voluntary and free of charge: no incentives (monetary or other) were provided to the participating teachers or schools. The PD program's after-school meetings included short lectures and presentations in relation to the three components. In addition, during almost all meetings, the participants were asked to work on

assignments “on the spot”. These assignments were individual assignments, assignments requiring collaboration between colleagues from the same school, and assignments requiring collaboration between colleagues from different schools. Four of the nine after-school meetings were set up as general gatherings in which all participants met in a convention center. The remaining five meetings were held at the individual schools (in a few cases, the participating staff members of two or three schools joined together in one meeting). In these meetings held at the school level, we focused on the performance of the teachers’ own students and provided the teachers with concrete suggestions relating to their own instructional behavior. In sum, the features of our program met the suggestions of the PD literature concerning a program’s *intensity* and *format*, as well as the recommended *collective participation of staff members from the same school* (e.g., Desimone, 2009; Garet et al., 2001).

The first and second component of the program were in line with recent developments in Dutch educational policy, which had become more oriented toward working with standards (i.e., performance goals based on performance categories) and using performance data to improve teaching (Ministry of Education, 2009; 2011). In the meetings on data use (the second component), teachers received training in using the student monitoring systems and they received assistance in interpreting the systems’ outcomes. The training sessions contained both easy as well as more complex assignments - including exemption from the easiest assignments -, in the acknowledgement of the differences between participants’ familiarity in working with student monitoring systems. Furthermore, for the meetings on instructional practices and key concepts in reading comprehension (the third component), we used anecdotes of the teachers’ own classroom practice which we had collected during observations - as part of the PD program - to better suit teachers’ knowledge and skills. Directly after the observations, the teachers received constructive feedback on their implementation of several instructional practices. Summarizing, the features mentioned above met the PD literature’s recommendations with respect to *congruence with national policy, active learning* of teachers and *congruence with teachers’ prior knowledge*, as well as *focus on content* (Desimone, 2009; Garet et al., 2001; Wayne et al., 2008; Yoon et al., 2007).

1.4 Theory of action and general research question

The aim of the PD program was to help teachers make their instruction in reading comprehension more goal-oriented and differentiated (as individual performance goals are defined and progress toward these goals is monitored), and to make instruction more focused and clear (by, among other things, the implementation of modeling and Direct Instruction and targeting teachers’ knowledge in this subject area). Through this assumed improvement in instruction, students’ reading comprehension performance was expected to improve.

In Figure 1, the theory of action is provided. To help explicate the role of each component in the integrated PD program, we connected these components to questions (c.f. Hattie & Timperley, 2007). The first component of setting standards and performance goals for every student is represented in the question: *Where am I going?* The second component on applying formative assessment and data use is represented in the question: *How am I going?* The third component on relevant instructional skills and (content and curriculum) knowledge in reading comprehension is represented in the question: *How can I improve how I am going?* In our theory of action, a distinction is made between the *theory of teacher change* in which the content of the PD is linked to change in teachers' practice, and the *theory of instruction* in which the changed practice is linked to change in students' attainment (see Wayne et al., 2008).

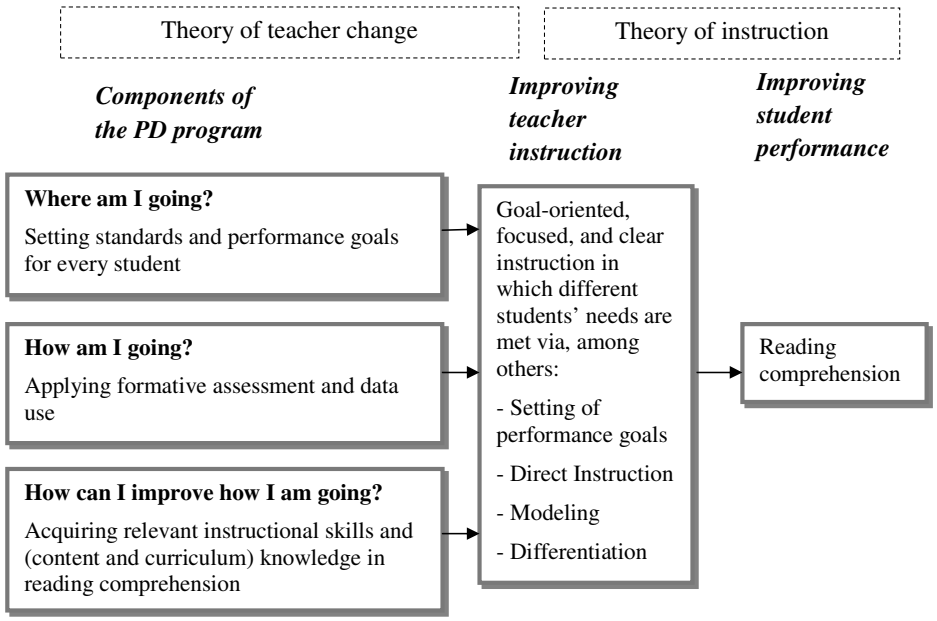


Figure 1. Theory of action

The goal of this dissertation is to evaluate the effectiveness of the multicomponent teacher PD program. The following research question is addressed: *Does students' reading comprehension improve after teachers have followed a multicomponent professional development program targeting goals, data use, and instruction, and can we find further empirical evidence for the assumptions underlying this program?*

1.5 Overview of the dissertation

An overview of the dissertation is provided in this section. In Chapter 2, we focus on the effect of the PD program on students' results, as the improvement of student reading comprehension was the main aim of the program. In Chapters 3 and 4, we concentrate on the performance goals which have been set by the participating teachers and which have played an important role throughout the program. More specifically, in Chapter 3 we study the topic of validity with respect to the performance categories on which the performance goals are based. In Chapter 4, we study the relation between the teacher-set performance goals and students' achievement. In Chapter 5, we investigate the participating teachers' implementation of a) Direct Instruction, b) modeling, and c) differentiation. Chapter 6 is the general conclusion and discussion of this dissertation, in which the findings and conclusions of the previous chapters are summarized. Furthermore, the limitations of the studies reported in the dissertation, several directions for future research, and the implications for practice are discussed in this chapter. In the Appendices of the dissertation one can find, among other things, a detailed description on the way the PD program has been conducted and information on the Dutch educational context in relation to the three components. In addition, lessons learned from the pilot study - conducted in the school year of 2010-2011 to help refine the program's design and materials - are discussed here as well.

The chapters in this dissertation are written in such a way that they can be read independently. Consequently, there will be overlap in the description of the background and content of the PD program.

2. The effect of the professional development program on reading comprehension

Abstract: In this chapter, we investigated whether student reading comprehension could be improved with help of a teacher Professional Development (PD) program targeting goals, data use, and instruction. A pretest posttest control group design was used to examine the effect of the PD program on second- and third-grade student achievement. Applying propensity score matching, 35 groups in the experimental condition were matched to 35 control groups. Students in the experimental condition (n = 420) scored significantly higher on a standardized assessment than students in the control condition (n = 399), with an effect size of d = .37. No differential effects of the PD program were found in relation to initial reading performance or grade. We checked for the robustness of these results using different model specifications, and found similar albeit smaller effect sizes for the effect of the PD program on student achievement (d = .29, d = .30, and d = .31, respectively).

2.1 Introduction

As reading constitutes an important basis for learning, working, and living, a common goal of primary schools is to equip students with sufficient reading skills (Kirsch, 2002; Reis et al., 2011; Snow et al., 1998). Currently, there are concerns on the reading results of Dutch students in primary school (e.g., Ministry of Education, 2008; 2010), particularly the degree to which struggling, poorly performing readers are prepared for later schooling and participation in society (Inspectorate of Education, 2007; 2010b). The assessment results on both national and international tests have been considered to be unsatisfactory, which has led to achievement concerns on the part of policymakers and the general public. For example, although the scores on the 2011 international PIRLS assessment (targeting fourth-grade reading) indicate that, comparatively speaking, students in the Netherlands perform rather well, the average achievement of the Dutch students is significantly lower than in 2001 (Meelissen et al., 2012). Furthermore, on the Dutch periodical assessment of educational achievement (known as PPON), 30 percent of the third-grade students read at a level which, according to reading experts and teachers, should be attainable for 75 percent (Van Berkel et al., 2007). These insufficient results have been ascribed to various causes, among which the lack of clear performance goals for teachers and schools to aim for in their teaching (Council of Education, 2007; Inspectorate of Education, 2011; Ministry of Education, 2010), the quality of teachers' reading comprehension instruction which could be improved in terms of explicitness (Aarnoutse & Weterings, 1995; de Jager, Reezigt, & Creemers, 2002; Van Elsäcker, 2002) and differentiation (Inspectorate of Education, 2012; Van Berkel et al., 2007; Van Elsäcker, 2002), and the hypothesized difficulty of teaching reading comprehension due to the complexity of the reading comprehension skills and inadequate curricular textbooks (Droop et al., 2012; Houtveen & Van de Grift, 2012; Stoeldraijer & Forrer, 2012). Hence, support has been gained for the view that the desired increase in achievement levels should be attained by targeting teacher instruction. To help teachers change their existing instructional routines, a multicomponent teacher Professional Development (PD) program was developed.

Teacher PD programs are a common tool for the implementation of performance improvement efforts (Bishop, Berryman, Wearmouth, & Peter, 2012; Hill, 2007; OECD, 2005) as there are substantial differences between teachers with respect to their ability to produce performance gains in their students (Nye et al., 2004). Researchers in the area of reading have emphasized the importance of schools and teachers to operate on a "what works" basis, in the context of which scientific evidence is used to improve the instructional practice and student performance (e.g., Armbruster et al., 2010; Collins Block & Lacina, 2009; National Reading Panel, 2000; Pressley, 1998). The call to use scientific evidence as a basis for the adoption of programs and practices not only applies to the subject area of reading, but to the whole field of education (see Slavin, 2008). In our PD program, the participating teachers were supported in

improving their practice with help of a three-component program: 1) setting standards and performance goals for every student, 2) applying formative assessment and data use, and 3) acquiring relevant instructional skills and (content and curriculum) knowledge in reading comprehension. All three components have shown to be positively related to student performance. We will describe the empirical evidence for each of these components[1] in detail further on. As the long-term effects of early acquired literacy skills have been documented in the literature (e.g., Bodovski & Youn, 2011; Snow, Burns, & Griffin, 1998) and the most striking results of interventions have been found for junior year groups (Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988), our PD intervention program specifically targeted teachers of the second and third grades (student age: approximately 7 to 9 years old). In the Netherlands, reading comprehension is a separate subject in the curriculum from the second grade onward.

In this chapter, we concentrate on the effect of teachers' participation in the PD program on student achievement as the improvement of students' reading comprehension was the general aim of the teacher PD program. As the three components had been integrated into one presumably synergetic package, we are interested in *molar causation* (Shadish, Cook, & Campbell, 2002, p. 509), being the overall relationship between the integral treatment package and its effects (rather than identifying the effectiveness of the separate components within the program). Below, we will elaborate on the theoretical background for the three components and specifications of the program will be provided in relation to each of these components. After this, additional general information regarding the set-up of the PD program is provided.

2.2 Theoretical framework

2.2.1. Component 1: Setting standards and performance goals for every student

Our first component pertained to setting goals. Working with goals has generally been proven to be effective for enhancing performance. Setting goals leads to a clearer notion of how desired outcomes can be attained, and it directs the focus toward the attainment of these desired outcomes (Fuchs et al., 1985; Fuchs, Fuchs, & Hamlett, 1989; Locke & Latham, 1990; 2002). When working with performance goals, these goals should be defined at a level that challenges teachers and their students, as ambitious goals are associated with higher outcomes (Locke & Latham, 1990; 2002).

[1] Not all components are associated with studies in which the participants were randomly assigned to conditions; a necessary criterion in order to speak of evidence for causation, see Borman et al. (2007). Many fields in education have as yet not been investigated using such experimental designs (see also Slavin, 2008). The components used in the current study were selected on the basis of evidence demonstrated in multiple studies and meta-analyses where a positive relationship between the issue under study and student achievement was found. Future research incorporating random assignment to conditions would be considered a valuable continuation of the research in these areas.

As part of the PD program, teachers were asked to set a performance goal for each of their students pertaining to the end of the school year. The goals were formulated by selecting one of five *performance categories* (labeled below minimum, minimum, basic, proficient, and advanced) in acknowledgement of the differences in students' capabilities. The performance categories were defined by the participating teachers in an early stage of the PD program with help of a *standard setting procedure*. More information on both the standard setting procedure and the setting of performance goals is provided in Chapters 3 and 4 of this dissertation.

2.2.2. Component 2: Applying formative assessment and data use

In order to help the participating teachers set appropriate goals and to help teachers attain these goals, it was important that they based their instructional decisions on assessment results (e.g., Guskey, 2002). Using student performance data to adapt one's teaching to meet students' needs is known as *formative assessment* (Black & Wiliam, 1998a; 1998b; Herman et al., 2010). In their meta-analysis, Black and Wiliam (1998b) conclude that there is "a body of firm evidence that formative assessment is an essential component of classroom work and that its development can raise (…) achievement" (p.148). Also other studies have shown that schools and districts applying a "data-driven" way of teaching can result in increased student performance levels (e.g., Carlson et al., 2011; Slavin et al., 2013). In the Netherlands, the progress of students is monitored with the use of *student monitoring systems*. Yet teachers are found to use these systems to a rather limited extent for the purpose of analyzing problems and, subsequently, adapting instruction. Furthermore, teachers who do use the student monitoring systems for these purposes are often unaware of the possibilities for more sophisticated analyses (Ledoux et al., 2009; Meijer & Ledoux, 2011; Schildkamp & Kuiper, 2010; van der Kleij & Eggen, 2013).

During the PD program, teachers received training in the use of the student monitoring system and in the interpretation of student performance data. Teachers were stimulated to work in reflective cycles of goal setting, planning, teaching, and monitoring. In the literature on data use and data-based decision making, the concept of performance data not only pertains to the assessment results on standardized tests, but also, for example, to student work or teacher observations of how the students function in class (Lai & Schildkamp, 2013). An important prerequisite, however, is that teachers adjust their practices accordingly after analyzing the data, a step which is not always guaranteed (Goertz, Olah & Riggan, 2009 in Carlson et al., 2011). Analyzing data "(…) is not enough to produce gains in achievement. Schools must actually take action to change teaching and learning" (Slavin et al., 2013, p. 390). The third component of our PD program focused on how to take action after analyzing the data.

2.2.3. Component 3: Knowledge and instruction for reading comprehension

After setting the goals and identifying the progress made toward them, it was important to help the teachers attain their own objectives by ensuring that they were sufficiently equipped with the most relevant instructional skills and knowledge about reading comprehension development. In order to advance students' reading proficiency, it is important that instruction is focused and clear (e.g., Andreassen & Braten, 2011; Verhoeven, 1991). Yet in the Netherlands, little explicit instruction is given during lessons in reading comprehension (Aarnoutse & Weterings, 1995; Van Elsäcker, 2002) and teachers hardly differentiate between students (Van Berkel et al., 2007; Van Elsäcker, 2002). It has been hypothesized, although not empirically researched, that primary school teachers in the Netherlands find reading comprehension a difficult subject to teach due to the complexity of the reading comprehension skills and the inadequacy of the curricular textbooks used in the Netherlands (Droop et al., 2012; Houtveen & Van de Grift, 2012; Stoeldraijer & Forrer, 2012).

In the PD program, two evidence-based instructional practices were discussed to help make instruction more focused and clear. These practices were *Direct Instruction*, being a teacher-centered model for instruction focused on the content and structure of a lesson (Borman et al., 2003; Muijs & Reynolds, 2011) and *modeling*, which is an instructional technique in which the teacher demonstrates how to apply a reading strategy or solve a problem by thinking aloud and linking the solution to skills or knowledge that the students already possess (Fisher et al., 2008; National Reading Panel, 2000). Modeling in combination with Direct Instruction has been identified as an effective procedure to help struggling learners and to remediate learning disabilities, as found in the meta-analyses of Swanson and Hoskyn (1998). Furthermore, we discussed important determinants of reading performance and key concepts in the second- and third-grade reading comprehension curriculum. More information is provided in Chapter 5 of this dissertation.

2.2.4. General information on the set-up of the PD program

Second- and third-grade teachers participated in our multicomponent program. In addition, the school's principal and the internal support coordinator participated, as their support was essential in facilitating the realization of change (e.g., Fullan, 2001). Throughout the school year, the time investment of the teachers was scheduled for 40 hours, including attending after-school meetings (nine in total; duration 1.5 to 2.5 hours per meeting) and completing homework assignments. Participation was voluntary and free of charge: no incentives (monetary or other) were provided to the participating teachers or schools.

2.2.5. *The current study*

As the improvement of students' reading comprehension was the main focus of the PD program, we concentrate on the effect of the program on student achievement. In addition, as some of the instructional practices are known to be particularly beneficial to struggling readers and the most promising results of interventions are found in the junior grades, we are interested in the relation between the effect of the program and students' initial reading performance as well as students' grade. In this chapter, the following research questions are addressed:

1) Does students' reading comprehension improve after teachers have followed a multicomponent Professional Development program targeting goals, data use, and instruction?

2) Does the effect of the program on students' reading comprehension depend on students' initial performance?

3) Does the effect of the program on students' reading comprehension depend on students' grade?

2.3 Method

To study the effect of the teacher PD on reading comprehension achievement, a quasi-experimental pretest posttest control group design was used.

2.3.1. *Participants*

Nineteen schools in the northern part of the Netherlands participated in the PD program. In total, 33 teachers took part in the program. These teachers taught 33 classes, of which 10 classes were multi-grade classrooms which contained both a second- and a third-grade year group. Thus, 43 groups of second- and third-grade students (containing 20 second-grade and 23 third-grade groups) were taught by the participating teachers. For the remainder of this chapter, we will refer to these groups rather than to teachers' classes as the matching procedure was conducted at the level of the group (i.e., the grade) rather than the class.

2.3.2. *Design and construction of the control group*

Our study formed part of a larger conglomerate of teacher PD intervention studies. This conglomerate was used to construct a suitable control condition. In total, over 90 Dutch primary schools participated in one (or, in a few cases, two) of five different teacher PD programs offered in the whole series of studies. The other PD programs targeted similar topics such as data use or standard setting. To promote the participation of the schools, they were given the opportunity to select the PD program of their choice. Each of the intervention studies targeted specific grades in

primary school: our study focused on the second and third grades. In order to construct a control condition for our study, we focused on the second- and third-grade groups from schools that had no intervention in these grades[2], forming a pool of possible control groups. Ultimately, a number of 80 groups, containing 56 second-grade groups and 24 third-grade groups[3], were identified as possible controls.

From this pool, we selected those groups that were the most similar to the groups in our experimental condition. Because we intended to take a number of group level characteristics into account, we applied the propensity score matching approach (Kelcey, 2011; Rosenbaum & Rubin, 1984). In this approach, the relevant group level variables are combined into one score which is an estimate of the probability of a group participating in the program. This score, with a value between 0 and 1 and estimated through logistic regression, is called the *propensity score*. The groups that were taught by teachers participating in the PD program were matched to the groups from the pool of possible controls on the basis of this score. An important feature of the propensity score matching approach is that the matches can differ in relation to the exact values of the variables used to estimate the propensity score. For example, average reading performance was one of the variables used to estimate the propensity score, and the average performance of a group of students in the experimental condition might differ from the average performance of its match in the control condition. Such a difference would not be problematic as the propensity score matching approach ensures that, overall, the variables used in the construction of the propensity score are sufficiently balanced when comparing the experimental and control condition (Rosenbaum, 2009).

After the propensity score was estimated for each group, several decisions had to be made regarding how to match these groups as there were different matching algorithms to choose from. Each matching algorithm has its own advantages and disadvantages - for more information, see Caliendo & Kopeinig (2008) and Heinrich, Maffioli, & Vazquez (2010). “It should be clear that there is no ‘winner’ for all situations (…) Pragmatically, it seems sensible to try a number of approaches. Should they give similar results, the choice may be unimportant” (Caliendo & Kopeinig, 2008, p. 44-45). We tested several options and found robust results in terms of balance among the variables used to estimate the propensity score as well as for the estimates of the effect of our program on achievement.

[2] Our PD program was the only one explicitly including the school principal and internal support coordinator (in addition to the participating teachers). This is why for the other PD programs as part of the larger series of intervention studies, we did not expect any ‘contamination’ of the second- and third-grade teachers via the school principal or internal support coordinator.

[3] This relatively small number of third-grade classes can be explained by the use of the multisubject test at the end of third grade to obtain an end-of-the-school-year measurement of the reading comprehension test: this test is detailed in the instruments-section. A precondition for inclusion of a third-grade class in the pool of possible controls was that this multisubject test had been administered.

In the results section of this chapter, we describe in detail the outcomes of a *caliper matching* algorithm we used. In caliper matching, a maximum propensity score distance can be specified between the propensity score of the class in the experimental condition and the propensity score of the class in the control condition. By applying this matching algorithm, we thus aimed to avoid the risk of selecting poor matches. The 0.2 caliper without replacement method (the "without replacement method" indicating that groups were matched on a one-to-one basis) yielded the most preferred results with respect to matching quality and sample size, as it resulted in the largest number of groups and students included in both the experimental and the control condition while creating a similar experimental and control condition. As such, the outcomes of the 0.2 caliper matching are presented in the results section below.

Next, we calculated the effect of the PD program on achievement using the student assessment results of the groups which had been selected via this matching method. To check for the robustness of our findings, we gathered the results of a 0.2 caliper with replacement: "with replacement" meaning that one group can be a match for multiple groups (this is detailed further on). In addition, we gathered the results of a 0.1 caliper without replacement, thus having a smaller maximum propensity score distance between the matched groups. The outcomes of these checks for robustness are presented in the results section as well.

2.3.3. Instruments and variables

Since the intervention was conducted at the group/teacher level[4], we wanted to match at this level. The data collected on the instruments discussed below were used for the estimation of the propensity score in order to identify suitable matches. The results on these instruments were also used to investigate the effect of the program. An important note pertains to the fact that the matching is done on the group level, whereas the prediction of student results – to identify the effect of the program – is done on the student level.

2.3.3.1 Instruments

Reading comprehension assessment: The Cito standardized reading comprehension assessments, developed by the Netherlands Institute for Educational Measurement, were used to measure the students' reading comprehension skills. The Cito standardized assessments form part of the assessment system most widely used in the Netherlands (LOVS); this LOVS system has been employed in approximately 85 percent of the Dutch primary schools (Inspectorate of

[4] Even though the group-level is not the same as the classroom level (as some groups were part of a multi-grade class), we refer to the group/teacher level for sake of simplicity as the results were found to be similar when higher levels of nesting were accounted for. More complex hierarchical models, including students nested in groups, nested in classrooms/teachers, nested in schools, were fitted to account for this structure of the data. The results of these models were almost the same as those of the model reported in this chapter; the effect of the program remained significant and had a similar size.

Education, 2010b). In the Cito reading comprehension tests, students are asked to read several texts and answer multiple-choice questions referring to the word, the sentence, and the text levels. These tests are administered from grade one to grade six. Both the validity and reliability of these tests have been considered sufficient: their reliability is above 0.89 (Cronbach's alpha) for the grades under study (Feenstra, Kleintjes, Kamphuis, & Krom, 2010). The tests have been approved by the Dutch National Committee of Tests and Testing, responsible for the review of tests (COTAN). The reading comprehension assessment results are registered on a continuous scale which ranges from -87 to +147 (end of grade 1 to mid-term grade 6). The negative symbol (-) for a large part of the assessment scale should not be interpreted as having a negative connotation; -87 is simply the (arbitrary) starting point of this scale.

The results of the June 2011 assessment[5] were used as one of the variables to estimate the propensity score for each group of students. These results were also used as a covariate in the prediction of students' individual results. For this latter analysis, the June 2011-reading comprehension results are referred to as the pretest. The June 2012-results were used to establish the effect of the program on students' achievement, and are referred to as the posttest. For most subject areas in the Cito LOVS standardized assessment system, assessments are conducted in each grade in January and in June. However, the reading comprehension assessment has a slightly different timing throughout the grades in primary school: it is conducted in June in the first grade, in both January and June in the second grade, and only in January from third grade onward. To obtain an "end-of-the-school year" result for reading comprehension in the third grade, we used the reading comprehension items of an additional multisubject standardized test (also developed by the Netherlands Institute for Educational Measurement), which is conducted in June. Its results are registered on the same scale as those of the regular reading comprehension assessments.

Mathematics assessment: The Cito standardized mathematics assessments were used as we wanted to incorporate students' prior math performances in the estimation of the propensity score as well as in the analyses on the effect of the program. By including students' mathematics results, we aimed to include a proxy for general academic ability. The mathematics test (also part of the Cito LOVS assessment system) is similarly approved by the Dutch National Committee of Tests and Testing (COTAN). Both the test's validity and reliability have been considered sufficient (Janssen, Verhulst, Engelen, & Scheltens, 2010), its reliability is above 0.91 (Cronbach's alpha). The mathematics results are registered on a continuous assessment scale which ranges from 0 to 169. The June 2011 grade-specific mathematics assessment results were used in our study.

[5] These results were collected at the end of grades one and two for the second- and third-grade students in our dataset.

2.3.3.2 Variables used for the estimation of the propensity score

We used seven group-level characteristics to estimate each group's propensity score via logistic regression. These seven variables were 1) grade (which was a dummy-coded variable with second grade as the reference category), 2) number of second- or third-grade students in the group, 3) percentage of girls in the group (girls generally outperform boys in reading, e.g., Bond, Dykstra, Clymer, & Summers, 1997), 4) the groups' average performance on the pretest, 5) the groups' standard deviation of performance on the pretest (indicating the level of heterogeneity in class), 6) the groups' average performance on the mathematics test, and 7) the groups' standard deviation of performance on the mathematics test.

2.3.3.3 Variables used for the analyses of the effect of the PD program on achievement

After estimating the propensity score and using the outcomes to identify a suitable control condition, we investigated whether students' posttest results (the dependent variable) could be predicted by participation in the PD program (the independent variable) with help of regression analyses. In these analyses, we controlled for the following covariates: a) students' sex (using a dummy-coded variable with boys as the reference category), b) students' grade (using a dummy-coded variable with second-grade as the reference category), c) students' performance on the pretest (grand-mean centered to facilitate its interpretation), and d) students' performance on the mathematics assessment (grand-mean centered to facilitate its interpretation). In a preliminary stage of our analyses, we had discovered that the relationship between the pre- and the posttest for reading comprehension was not directly linear, but could better be described by including polynomial terms. Including quadratic and cubic transformations of the reading comprehension pretest significantly improved the fit of our models, and these transformations were therefore included. The propensity score itself was not a significant predictor of student performance and was therefore excluded from these analyses.

2.3.4. Analyses

The *PSmatching* program, an *R* Plugin written as a so-called custom dialog in SPSS, was used to conduct the task of estimating the propensity scores and matching the conditions (Thoemmes, 2012). We investigated the quality of the matching procedure using an independent samples *t*-test, which is a common way to assess the quality of matching procedures. After matching, no significant differences should be found between the means of the variables used to estimate the propensity score (Caliendo & Kopeinig, 2008).

In order to assess the effect of the PD program on achievement, a multilevel regression analysis was performed with the help of *MLwiN* software (Rasbash, Browne, Healy, Cameron, & Charlton, 2011), with students (level 1) nested in groups (level 2). In this analysis, it was investigated whether the performance on the standardized reading comprehension posttest could

significantly be predicted by participation in the program while controlling for the aforementioned covariates.

2.4 Results

2.4.1. The quality of the matching results

The results of the propensity score matching approach are discussed first. After using the 0.2 caliper matching algorithm without replacement, it was found that not all groups in the experimental condition could be provided with a match. Of the 43 groups, 35 treated groups were matched to 35 untreated groups. Both conditions contained 19 second-grade groups and 16 third-grade groups. The groups in the experimental condition which could not be matched were characterized by relatively high propensity scores, while the unmatched groups in the pool of possible controls had mostly low propensity scores. In Table 1, we present the descriptive statistics of the variables used to estimate the propensity scores, both before and after matching. The statistical significance of the differences between the experimental and the control condition, tested using the independent samples *t*-test, are – if present – denoted by an asterisk in column b (in which the experimental and the control condition are compared *before* matching) and column d (in which the experimental and the control condition are compared *after* matching).

Table 1

Group Characteristics in Experimental and Control Condition before and after Matching (Means of Means)

	Before matching				After matching			
	Experimental (a) (n = 43)		Control (b) (n = 80)		Experimental (c) (n = 35)		Control (d) (n = 35)	
Variables	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*
Grade	0.54	(0.51)	0.30*	(0.46)	0.46	(0.51)	0.46	(0.51)
Number of students	10.49	(6.98)	14.78*	(7.24)	12.00	(6.83)	11.40	(5.15)
Percentage of girls	0.50	(0.21)	0.50	(0.14)	0.48	(0.16)	0.48	(0.14)
Average performance on pretest	12.39	(12.99)	8.12	(11.43)	11.11	(13.56)	13.17	(11.27)
SD of performance on pretest	13.42	(6.24)	13.11	(3.38)	13.70	(5.40)	13.24	(3.87)
Average performance on math test	55.95	(12.71)	51.99	(11.42)	54.49	(12.52)	56.11	(11.82)
SD of performance on math test	13.07	(4.84)	13.15	(3.69)	14.05	(4.36)	13.19	(4.37)
Propensity score	0.43	(0.17)	0.30*	(0.15)	0.39	(0.17)	0.39	(0.15)

* $p < .05$.

As can be seen from the results presented in Table 1, the groups in the experimental condition and in the pool of possible controls differed significantly with respect to three variables (i.e., grade, number of students, and propensity score) prior to matching. After matching, the group level variables were distributed in a much more balanced way and the differences between the conditions were no longer significant.

2.4.2. The effect of the program on students' reading achievement

After establishing an equivalent experimental and control condition, we assessed the effect of the PD program on student achievement using the multilevel regression analyses. The dataset used for these analyses consisted of the results of 819 students in total. The results of 420 students (51 percent) in the experimental condition were compared to those of 399 students in the control condition. In Table 2, the descriptives statistics of the variables used in the regression analyses are presented. Tables 1 and 2 differ in that the first refers to the group level and the second to the student level.

Table 2

Student Characteristics in Experimental and Control Condition

	Experimental		Control	
Variables	*M (SD)*	*n* (%)	*M (SD)*	*n* (%)
Grade 2		231 (55)		210 (53)
Grade 3		189 (45)		189 (47)
Girls		201 (48)		193 (48)
Boys		219 (52)		206 (52)
Math performance	54.79 (17.29)		56.27 (18.03)	
Pretest	10.82 (18.80)		13.82 (16.44)	
Posttest	27.24 (16.75)		25.91 (15.09)	

The results presented in Table 2 show that the variables grade and sex were similarly distributed across the conditions. The average performance on the pretest was lower in the experimental group than in the control group. As a result of the lower performance results of the students in the experimental condition, we might be at risk of mistaking a *regression-to-the-mean effect* (see Cozby, 2003) for a positive effect of the intervention on the students' achievement. We checked this difference between conditions while taking the multilevel structure of the data (students nested in groups) into account, and found that the difference between the conditions on the pretest was not significant ($t = -.66$, $p > .05$). The same was done

for the mathematics' results as the average performance on the mathematics test was slightly lower in the experimental condition as well. Again, the conditions were found not to differ significantly (t = -.49, p > .05).

Table 3 reports the results of the multilevel analyses, where the program's effect on reading achievement was estimated while controlling for the relevant covariates. The first model, called the start model, contained the intercept and the covariates at the student and the group level but did not include the variable indicating whether the teacher had participated in the PD program. In the second model, called the main effect model, this variable was included. In this way, we could analyze whether or not this variable added value when predicting student achievement. In the third model, called the interaction model, we investigated differential effects as we added an interaction term between participation in the program and students' pretest performance, as well as an interaction term between participation in the program and students' grade. All models presented in Table 3 contain unstandardized coefficients.

Table 3

Multilevel Models Predicting Achievement in Reading Comprehension

	Models					
	Start model		Main effect model		Interaction model	
Predictors	Coeff.	*SE*	Coeff.	*SE*	Coeff.	*SE*
Fixed Part						
Intercept	19.16*	0.92	17.21*	1.03	17.23*	1.17
Grade 3	-3.82*	1.30	-3.74*	1.23	-3.10	1.71
Girls	2.29*	0.75	2.29*	0.74	2.24*	0.75
Math performance	0.17*	0.03	0.16*	0.03	0.17*	0.03
Pretest	0.67*	0.04	0.67*	0.04	0.65*	0.05
Pretest^2	0.00165*	0.00067	0.00155*	0.00067	0.00166*	0.00069
Pretest^3	-0.00006*	0.00001	-0.00006*	0.00001	-0.00006*	0.00001
Participation in PD program			3.78*	1.10	3.85*	1.53
Participation x pretest					0.04	0.05
Participation x grade					-1.27	2.40
Random Part						
Variance at class level	14.03	4.00	11.21	3.53	11.17	3.51
Variance at student level	101.97	5.26	101.74	5.24	101.93	5.25
Deviance	6168.64		6157.17		6156.48	
No. of groups	70		70		70	
No. of students	819		819		819	

* $p < .05$.

When comparing the start model to the main effects model, one can see that participation in the PD program is related to a significant higher performance on the posttest. Inclusion of this variable fairly increased the fit of the model: the deviance decreased by 11.47, which is a significant improvement (the critical value in a chi-square distribution with $df = 1$ is 3.84 for $p = .05$, as the models differ in 1 parameter). When comparing the main effects model to the interaction model, the interactions were found to be non-significant and the model fit had not improved. In other words, the positive effect of the program on students' achievement was found to apply irrespective of students' initial performance on the pretest, and the PD program was equally effective for second- and third-grade students.

For the effect of the PD program on student achievement, we observed an effect size of $d = .37$ (calculated by dividing its regression coefficient by the square root of the unexplained variance at the student level, using the coefficient in the main effects model[6]), 90% CI [$d = .20$; $d = .55$]. According to Cohen's interpretation (1988)[7], a value of $d = .37$ is a small to medium effect.

2.4.3. Checks for robustness

To check the robustness of our results, various methods were used. First, we investigated whether participation in the program remained a significant predictor of achievement when separately modeling the outlying residuals at both the student and the group level. Outlying cases might have an "undue high influence on the results of the statistical analysis" (Snijders & Bosker, 1999, p. 128), and we wanted to ensure that the positive effect of participation on achievement - as identified in the current study - was not caused by influential outliers. In identifying the outliers, we made use of z-scores[8]. In the outlier model we fitted, participation in the PD program was a significant positive predictor of student achievement with an effect size of $d = .29$, 90% CI [$d = .13$; $d = .45$].

In addition, we checked the robustness of our results by using different propensity score matching methods. Below we will list our findings using a 0.2 caliper with replacement and a 0.1

[6] This is an application of Cohen's (1988) formula of $d = \{\bar{x}(exp) - \bar{x}(control)\}/\sigma$ to a multilevel setting, for which we are interested in the variation within groups (i.e., level one).

[7] Cohen (1988) provides the following guideline for the interpretation of effect sizes: $d = 0.2$ is considered to be a small effect, $d = 0.5$ a medium effect and $d = 0.8$ a large effect.

[8] At the student level, outliers were defined as values with standardized scores lower than $z = -3.29$ or larger than $z = 3.29$ (Tabachnick & Fidell, 2001). Three students could be flagged as having an outlying score after fitting the model, which was caused by an extremely high result on the posttest. For the standardized residuals at the classroom level, we used a stricter z-score criterion, as outlying cases at the classroom level may influence the model more substantially than outlying cases at the pupil level (Rasbash, Steele, Browne, & Goldstein, 2012). In total, 4 classes with a standardized residual below $z = -2$ or above $z = 2$ were modeled separately. In addition to using this z-score approach to outlier identification, we also checked the influence of outliers with help of a method proposed by Tukey (1977) which makes use of P25 (the first quartile), P75 (the third quartile), and the Inter-Quartile Range (IQR): here outliers have values below P25 – 1.5 x IQR and above P75 + 1.5 x IQR. This approach yielded a larger effect size for the program, with $d = .40$, 90% CI [0.23; 0.58].

caliper without replacement. When matching with replacement, the control groups could participate several times in the control condition (these control groups were given larger weights in the analyses of the effect of the program, as they were matches for several groups in the experimental condition). Using the 0.2 caliper with replacement method, 20 second-grade and 20 third-grade groups in the experimental condition were matched to 14 second-grade and 15 third-grade groups in the control condition. The differences between the experimental and the control condition for the variables used to construct the propensity score were non-significant (tested by applying the independent samples *t*-test). When using the assessment results of the students in the groups selected via this matching method, we found a similar effect size of $d = .30$, 90% CI [$d = .12$; $d = .48$] for the effect of the program on the students' reading achievements.

When applying the 0.1 caliper without replacement method (thus having a smaller maximum propensity score distance between the matched classes), 19 second-grade and 11 third-grade groups in the experimental condition were matched to 17 second-grade and 13 third-grade groups in the control condition: not all groups were thus matched to groups of same-grade students. This finding can be explained by the fact that grade was only one of the variables used in the matching procedure. By adopting a propensity score matching approach, an overall balance is attained for the variables used in this score's estimation. Some variables might then be more similarly distributed over the conditions than other variables. Again, we tested the equivalence of the experimental and control condition after matching with help of the independent samples *t*-test. None of the variables used to construct the propensity score – including the variable grade - differed significantly between the conditions. We found a rather comparable size of $d = .31$, 90% CI [$d = .13$; $d = .48$] for the effect of the program on student achievement via this method[9].

Summarizing, the three alternative methods used to check for the robustness of our original results yielded a similar positive outcome for the effect of the PD program on achievement, with slightly smaller but fairly comparable effect sizes.

2.5 Conclusion and discussion

In this chapter, we investigated whether reading comprehension performance could be improved with help of a teacher Professional Development (PD) program which had been developed as a response to the recent performance concerns in the Netherlands. Given the importance of the early acquired literacy skills, we specifically targeted second- and third-grade students in this improvement effort. The PD program was designed to foster student reading comprehension through teachers' application of a multicomponent package. Using the propensity score matching approach to construct an equivalent control condition from a larger pool of

[9] The exact results of the independent samples *t*-tests as well as the multilevel analyses of the program's effect using the 0.2 caliper with replacement and the 0.1 caliper without replacement datasets will be made available on request after contacting the first author.

possible controls, we found that students in the experimental condition performed significantly better than those in the control condition, with an effect size of $d = 0.37$; 90% CI [$d = .20$; $d = .55$]. We checked for the robustness of these results using different model specifications, and found similar albeit smaller effect sizes for the effect of the PD program on student achievement ($d = .29$, $d = .30$ and $d = .31$, respectively). According to Cohen's interpretation (1988), these effect sizes can be interpreted as small to medium effects.

Differential effects of the program on student achievement were investigated but these were non-significant. All students, irrespective of whether they were initially low or high achieving students or whether they were in second or third grade, appeared to have profited equally from their teachers' participation in the PD program. The PD program was not designed to target certain subgroups specifically; hence we did not expect such differential effects. Should the improvement of struggling readers be of prime interest - relevant in light of the performance concerns in the Netherlands pertaining to this specific group -, more intensive didactical practices (such as one-to-one instruction) are necessary.

Several limitations to the current study should be considered. In order to facilitate participation, schools could choose whether or not they wanted to participate in our PD program – they were not randomly assigned to the experimental or the control condition. In order to account for differences between these conditions we used the propensity score matching approach, but the number of variables used to estimate the propensity score as well as the number of groups in the pool of possible controls were relatively small. We originally wanted to include teacher characteristics (such as years of experience as well as more affective properties such as teachers' attitude toward data use) to obtain more detailed information on the level where the intervention was conducted (the teacher/classroom level). Yet the degree of non-response to a teacher questionnaire, one of the instruments used in the larger series of intervention studies, prohibited us from doing so. Nonetheless, all schools in the series of intervention studies participated because they wanted to improve their education through participation in PD programs targeting similar topics. All schools and teachers were aware of students' results being measured throughout the entire school. Therefore, relatively similar schools and teachers are considered to have taken part in both the experimental and the control condition. Nevertheless, a replication of this study in which schools are randomly assigned to conditions would complement the findings of the current study as they allow for stronger statements on the causation of the program's effects (i.e., there would be no threat of omitted variable bias).

Another consideration pertains to the intensity of the program and the fact that the participants knew that they were trained by researchers. In the case of a *Hawthorne effect* (Shadish et al., 2002) the positive effect of our PD program on student achievement could have been caused by the fact that participants improved their behavior simply because of the knowledge that they were being studied, and not because of the content of our program. Yet we find this Hawthorne explanation somewhat improbable given the complex nature of the reading

comprehension skill (e.g., Afflerbach, Pearson, & Paris, 2008). Not all teacher professional development programs targeting reading succeed in significantly improving student results (see the overview of studies in Yoon et al., 2007). In the review study of Yoon et al. (2007), the most promising results of teacher professional development are found for those studies including a focus on how students learn and how to assess student learning (although the number of studies included in this review was rather limited; only nine studies met the What Works Clearinghouse criteria). Successful stimulation of reading performance might require PD programs consisting of a combination of relevant components on content and data. This line of reasoning seems to be confirmed by the multicomponent *Success for All* program (Slavin et al., 1996) targeting the prevention of and early intervention in reading difficulties. In this program, among other things, teachers use a prescribed curriculum in order to provide high-quality instruction and students' progress is frequently monitored. After three years of continuously implementing this Success for All program, effect sizes between $d = .21$ and $d = .36$ were found in kindergarten to grade two across various reading performance measures (Borman et al., 2007). To attain reading performance improvement as we did in our study, teachers are assumed to have provided high quality instruction. For this, we consider the content of our program to have been essential; the implementation of the program by teachers and other assumptions underlying the program will be studied in the forthcoming chapters of this dissertation.

In the current chapter, we studied the effect of a multicomponent teacher PD program which was aimed at improving student reading comprehension. The significant higher reading results of students in the experimental condition lend support for the conclusion that the program was successful in attaining its aim.

3. Investigating the validity of cutscores

Abstract: Teacher-set performance goals played a key role in a teacher Professional Development (PD) program aimed at improving reading comprehension. The performance goals were formulated in terms of performance categories, and these categories had been established by the participants of the PD program with help of a standard setting procedure. In this procedure, the participants were asked to identify the boundaries of the performance categories through multiple rounds of standard setting. These boundaries are referred to as cutscores. According to the standard setting literature, the evaluation of the accuracy of these cutscores should be done by investigating the evidence for different types of validity. In the current chapter, the procedural validity of cutscores was studied with help of participants' feedback pertaining to a) the procedure's explicitness, b) the procedure's practicability, and c) the panelists' deliberateness. The internal validity was assessed through the investigation of d) the panelists' adaptations across rounds, e) the correspondence between cutscores and empirical performance data, and f) the interpanelist agreement. The results of the analyses indicated sufficient support for both types of validity.

3.1 Introduction

Working with goals is effective for enhancing performance as goals direct the attention toward the attainment of desired outcomes (e.g., Fuchs et al., 1985; Fuchs et al., 1989; Locke & Latham, 1990; 2002). This explanation is also used as a rationale for the implementation of standards in education: it is assumed that standards will improve instruction and subsequently student performance, as schools and teachers are provided with clear goals to aim for in their teaching (Lauer et al., 2005; Roeber, 1999). Educational performance standards are examples and definitions of what students have to know and do (Ravitch, 1995), and these performance expectations have been developed in such a way that they apply to all students - ranging from low to high achievers -, as different categories of proficiency (for instance basic, proficient, and advanced) are identified in the standards. Because of this expectation of performance improvement, standards-based education is currently employed in several countries, among which the United States, England, Germany, and Australia, in which attainment targets are formulated in terms of standards (OECD, 1995; Pant, Rupp, Tiffin-Richards, & Köller, 2009).

Aiming to improve the early reading performance of Dutch students following recent performance concerns (e.g., Ministry of Education, 2008; 2010) and acknowledging the widely established importance of early acquired literacy skills (Bodovski & Youn, 2011; Snow et al., 1998), we developed a teacher Professional Development (PD) program targeting second- and third-grade teachers (student age: 7 to 9 years old). As part of the program, teachers were asked to set performance goals for their students. The goals were formulated by selecting one of five *performance categories* (labeled below minimum, minimum, basic, proficient, and advanced), in acknowledgement of the differences between students' capabilities. In order to establish these performance categories, a so-called *standard setting procedure* was conducted during the PD program. Commonly these procedures are used to identify performance categories on, e.g., state-wide or national tests in countries working with standards (Cizek & Bunch, 2006). A specific feature of performance categories which are set in many standard setting procedures is that they pertain to test score intervals on assessment scales; in our PD program, the performance categories pertained to the scale of the standardized reading comprehension assessments. The benefit of working with goals that are established in terms of test scores on an assessment scale is that the attainment of these goals is easily established by conducting the assessments in class.

Standard setting procedures entail that participants discuss and reflect on their performance expectations for students with different capabilities. In many standard setting procedures (including the one conducted as part of our program), the panelists are asked to identify the boundaries or *cutoff points* of the successive performance categories. These cutoff points are referred to as *cutscores* or *cutoff scores*. After using a standard setting procedure to formulate cutscores, the *validity* of these cutscores should be evaluated (Pant et al., 2009). In standard setting, validity pertains to the degree in which the classifications of test results into performance

categories as well as the actions following these classifications are considered accurate (Cizek & Bunch, 2006). To help evaluate the validity of cutscores, evaluation criteria have been proposed in several standard setting guidelines (Cizek & Bunch, 2006; Hambleton & Pitoniak, 2006; Kane, Crooks, & Cohen, 1999; Norcini & Shea, 1997; Pant et al., 2009).

In this chapter, we investigate the degree to which the standard setting results of our program meet the evaluation criteria. During the PD program, the performance goals which are based on the performance categories play a key role, and it is important to evaluate the underlying assumption that the performance categories and associated cutscores are accurate. The current evaluation is interesting for other researchers in the area of standard setting as the number of empirical evaluations of standard setting procedures pertaining to the cutscores' validity is limited despite the availability of evaluation guidelines (also in McGinty, 2005).

3.2 Theoretical framework

3.2.1. Standard setting procedures

Usually, two types of standards are identified, namely a) content standards, and b) performance standards (Hambleton & Pitoniak, 2006). Content standards define what should be taught and what students should learn. Performance standards are examples and definitions of what students have to know and do, to demonstrate proficiency in the knowledge and skills framed by the content standards (Ravitch, 1995). The term *standard setting* pertains to the performance standards and is used to refer to the procedure in which cutoff scores are set on test result scales in order to establish the successive performance categories. These categories can be seen as being *criterion-referenced* categories, in which a student's test performance is classified into a particular performance category based on the test score, in contrast to *norm-referenced* categories in which performance is classified into a particular performance category on the basis of performance relative to other examinees and pre-specified percentages (Hambleton, 2001). For the development of performance categories with help of standard setting procedures, human judgment is required (Berk, 1986) and the standard setting procedures are heavily structured to facilitate this task (discussed in, for example, Cizek & Bunch, 2006).

Standard setting procedures are either *test-centered* approaches which involve judgments about test items (by identifying items which should answered correctly by students of a certain level) or *examinee-centered* approaches which involve judgments about candidates and/or looking at candidate work (Cohen, Kane, & Crooks, 1999). For either of these approaches, there are many types or methods of standard setting (e.g., Cizek & Bunch, 2006). Regardless of which type or method is selected, several common steps are required. The first step is that those responsible for standard setting must select appropriately qualified panelists (also called judges) to participate in the procedure. The standard setting panel should contain representatives of

different groups for whom the assessment results and the decisions following the attainment of certain categories are relevant. At the same time, it is important that the panelists have sufficient experience with the student population for which the performance categories are defined. This is because they are asked to envision students with different abilities and to identify the kind of performance that would be consistent with a student of a certain ability (Plake, 2008; also in Raymond & Reid, 2001).

When attending the actual standard setting meeting, panelists should receive an introduction on the purpose of standard setting. The goal of the procedure should be made clear to them from the start (Cizek & Bunch, 2006; Hambleton & Pitoniak, 2006). Furthermore, as panelists will be asked to identify multiple cutoff points as part of the standard setting procedure, clear descriptors should be provided to illustrate these different cutoff points. It is known that panelists find it difficult to properly interpret multiple cutoff points and to envision student performance which corresponds to these points (Poggio, 1984 in Berk, 1986; Plake, 2008). By providing them with descriptors that illustrate these cutoff points, a "common understanding of performance, that is, a frame of reference" (Hambleton & Pitoniak, 2006, p.455) is generated. Another aspect which is addressed relatively at the beginning of the standard setting meeting, is training in the use of the standard setting materials such as booklets or forms. During this training, the set-up of the different rounds should also be addressed and a clear account should be provided on what is expected of panelists during different standard setting rounds (e.g., in Deunk, van Kuijk, & Bosker, in press).

After the training, the panelists commence with the setting of cutscores. Standard setting procedures typically make use of various rounds to foster convergence of panelists' views, as it is the aim of such procedures to end up with cutscores that have been agreed upon by multiple experts (Hurtz & Auerbach, 2003; Karantonis & Sireci, 2006). The following three rounds are frequently used for standard setting purposes: in the first round, the panelists study the materials by themselves and individually identify the cutscores that they consider to be appropriate cutoff points for the different categories. In the second round of the procedure, the panelists are asked to engage in small group discussions. During this round, they explicate the grounds for their first-round cutscores. This discussion and sharing of information is expected to provide panelists with a more comprehensive view of factors that are relevant to student performance. The discussion round is assumed to help the participants to (re-)set their cutscores at a more accurate level, reflecting their views on desired student performance at different levels. When (re-)setting their scores, the groups are allowed to reach consensus but this is not mandatory. In the third round, panelists are provided with *performance information* (Hurtz & Auerbach, 2003) which is also referred to as normative information or consequence data (Busch & Jaeger, 1990; Hambleton & Pitoniak, 2006). This information shows how the current student population would perform in relation to the cutscores which have been set in the previous round. Presenting this information

may lead to the resetting of cutscores when, in the words of Hambleton and Pitoniak (2006): "consequence data are not consistent with panelists' experiences and sense of reasonableness" (p. 455), that is, when too many or too few students fall into certain categories. At the end of the procedure, the medians of the cutscores set during the last round are taken as the final cutscores. As the cutscores identify the boundaries of test score intervals, the associated performance categories are then established onto the scale of a certain assessment.

After the panelists have set their final cutscores at the end of the standard setting procedure, an evaluation of the standard setting procedure should be conducted by those responsible for the standard setting procedure. An important aspect of this evaluation is asking the participants whether they, among other things, were able to understand and take part in the different rounds of the procedure. Merely conducting a certain standard setting procedure does not guarantee that the final cutscores are valid; poorly implemented procedures can affect the accurateness of these scores. As standard setting is a complex task and high-stakes consequences can follow the attainment of particular performance categories - as is the case in the United States, ranging from licensure of an examinee to accreditation of an institute (Linn, 2000) -, it is important to evaluate the validity of the standard setting results (Deunk et al., in press; Hambleton & Pitoniak, 2006; Pant et al., 2009).

3.2.2. Validity

When focusing on the validity of standard setting results, one needs to distinguish between the performance standard and the cutscore. While the former pertains to the conceptual version of a desired level of competence, the latter pertains to the operational version of that desired level of competence (Kane, 1994). Validity thus entails a twofold assumption in this context: 1) the performance standard is appropriate given its intended use, and 2) the cutscore is an appropriate representation of the performance standard. Different types of evidence can be collected to validate these assumptions, namely 1) external, 2) procedural, and 3) internal evidence (Hambleton & Pitoniak, 2006). Other authors use a slightly different categorization (e.g., Pant et al., 2009), but the underlying content of these categories is rather similar.

External evidence is generally used to check the validity of the performance standard, i.e., the conceptual version of desired levels of competence. It focuses on the classification of performance into performance categories and the consequences that follow this classification. For example, it should be investigated whether the decisions following the attainment of certain categories are in line with broader policy goals for which the standards were developed (Hambleton & Pitoniak, 2006; Pant et al., 2009). In this chapter, we focus on the validity of the cutscores as it is the operational version of the desired level of competence which is of interest here. For this purpose, we will continue to the criteria for procedural evidence of validity, followed by the criteria for internal validity.

Procedural evidence of cutscore validity is obtained by analyzing the set-up, interpretation, and execution of the standard setting procedure. Poorly conducted procedures will result in less credible cutscores. Several criteria have been proposed in the literature[10] to help assess the cutscores' procedural validity. First of all, the procedure should be clearly explained; this criterion is referred to as the procedure's *explicitness.* Furthermore, the procedure should be *practicable* which is measured by evaluating the degree to which the standard setting method was implemented without great difficulty. In addition, panelists should be asked to evaluate their own *deliberateness*; whether they consider that their own cutscores have been set in a well-considered manner and have confidence in them. "[I]f the judges who developed the standards do not have confidence in it, it is not clear why anyone else should" (Kane, 1994, p. 443). Feedback from the panelists themselves is considered an important source of information (also in, Hambleton & Pitoniak, 2006; Kane, 1994; 2001).

Internal evidence of cutscore validity concerns the amount of variation in cutscores, and the following five criteria can be deduced from the literature[11]. The first criterion pertains to *panelists' adaptations across rounds*. If the panelists are taking in and synthesizing the information they are provided after each round, these scores should show some variability. In contrast, if panelists are staying with their initial cutscore round after round, this calls into question the effectiveness of the different rounds which aim to help panelists refine their ratings. Panelists can make a well-considered decision to stay with their initial cutscores, not changing cutscores is "not necessarily a sign of poor standard setting" (Hambleton & Pitoniak, 2006, p. 460). Yet variation in cutscores is considered a sign that the information is used whereas static cutscores do not indicate such use. A second criterion to evaluate the internal validity of cutscores pertains to the *correspondence between cutscores and empirical performance data*. If panelists are found to differ from empirical data to a very large extent, this calls into question whether the cutscores are realistic. A third criterion which is used to assess cutscores' internal validity is to identify whether the variation in cutscores decreases after several rounds of standard setting, indicating that panelists' cutscores converge. When the degree of variation remains stable across the rounds, this may indicate problems in the training of the participants or in the materials used. The agreement between panelists as indicated by the decrease in variation in cutscores is called *interpanelist agreement*. Ideally, it is advised to also compare cutscores across panels (the *across-panel consistency)* and across content areas or cognitive processes (*across-subject consistency*) but this

[10] Hambleton and Pitoniak (2006) provide the follow categorization of criteria to assess the cutscores' procedural validity, namely: a) explicitness, b) practicability, c) implementation of procedures, d) panelist's feedback and e) documentation. As these five criteria of procedural validity contained overlap in their content and how they could be measured, we combined them into three criteria.

[11] Hambleton and Pitoniak (2006) refer to the following internal validity criteria: 1) intrapanelist consistency-between steps, 2) intrapanelist consistency-within steps, 3) interpanelist consistency, 4) within method consistency or the reliability of cutscores, and 5) other measures of consistency. Again, due to overlap in content, we use a slightly different categorization.

is generally impractical and expensive (Pant et al., 2009). The criteria which can be used for the evaluation of the procedural and internal validity are summarized in Table 1.

Table 1

Overview of the Criteria for the Evaluation of the Procedural and Internal Evidence of Cutscores' Validity

Type of validity	Criteria	Explanation The degree to which …
Procedural	Explicitness	.... the participants consider the procedure to be clear.
	Practicability	… the participants consider the implementation of the procedure to be practicable.
	Deliberateness	… the participants consider their own cutscores to be well-considered.
Internal	Panelists' adaptations across rounds	… the participants revise their own cutscores across the rounds.
	Correspondence between cutscores and empirical performance data	… the cutscores concur with the performance data.
	Interpanelist agreement	… the participants come to an agreement, and there is a decrease in variation around cutscores.
	Across-panel consistency	… the cutscores are consistent across panels.
	Across-subject consistency	… the cutscores are consistent across subjects or cognitive processes.

An important note is that the evidence which can be collected with respect to these criteria is conditional: evidence of a well-implemented standard setting procedure and high agreement among panelists do not necessarily imply the validity of the resulting cutscores, whereas proof of procedural flaws and a low inter-rater agreement may point to a lack of validity (Kane, 1994). Nonetheless, by evaluating the degree to which these criteria are satisfied, one attains an indication of the cutscores' validity. Positive evaluation results are viewed as support for the cutscores' use.

3.2.3. The current study

After conducting a standard setting procedure, it is important to investigate the validity of the cutscores. Also in the context of our teacher PD program, it is important to evaluate the assumption that the cutscores and associated performance categories are valid. In this chapter, the

validity of cutscores is evaluated with help of criteria proposed in the standard setting evaluation literature. By reporting our results, we aim to provide an empirical example to make the guidelines on standard setting evaluations and validity issues more tangible. In the current study, we focus on the following research question: *To what extent do the results of the current standard setting procedure meet the criteria for the procedural and internal validity of cutscores?* With regard to the procedural validity, we target the criteria of a) explicitness, b) practicability, and c) deliberateness. With regard to the internal validity, we target the criteria of d) the panelists' adaptations across rounds, e) the correspondence between cutscores and empirical performance data, and f) the interpanelist agreement.

3.3 Method

For the investigation of the cutscores' procedural validity, a one-group posttest-only design was used as the participants' evaluations were collected at the end of the standard setting procedure. For the investigation of the cutscores' internal validity, panelists were asked to set cutscores during multiple rounds; here, the data have come from repeated measurements.

3.3.1. Participants

The standard setting procedure was conducted in the school year of 2011-2012 with participants from nineteen schools in the northern part of the Netherlands. In total, 67 panelists participated in the standard setting study. This sample consisted mostly of teachers (n = 46[12]; 69 percent) in addition to a number of school principals and internal support coordinators. In this sample, the vast majority of the participants (n = 56; 84 percent) were female. The average experience in teaching was 15.72 years (SD = 11.38).

The teachers set cutscores for the grade they were teaching at that time, i.e., being second or third grade. The school principals and internal support coordinators were allocated to one of the two grades at random. A slight majority of the participants set cutscores for second-grade reading comprehension (n= 39, being 58 percent); the other participants set cutscores for third-grade reading comprehension.

3.3.2. Bookmark standard setting procedure

In this study, the Bookmark method (Mitzel, Lewis, Patz, & Green, 2001) was used to set cutscores. This particular standard setting method is a popular method in the United States, often used for state assessment systems (Karantonis & Sireci, 2006) and for NAEP, the United States' national assessment system (Peterson, Schulz, & Engelhard, 2011). We used an adaptation of the

[12] This number of teachers is larger than the number mentioned in Chapters 1 and 2 of the dissertation. Several of the teachers in our program were teaching part-time. For a number of these teachers, their partner-colleagues (who did not provide reading comprehension instruction in the school year under study) attended the PD program's meetings.

Bookmark procedure as commonly used by the Netherlands Institute of Educational Measurement (Van der Schoot, 2009) with whom we cooperated in the preparation of this study. In the Bookmark procedure, an Ordered Item Booklet (OIB) is used in which items are ordered to increase in difficulty with help of Item Response Theory (IRT).

The goal of the procedure was to create five performance categories (below minimum, minimum, basic, proficient, and advanced) by asking participants to place four bookmarks and thus identifying four cutoff points. The bookmark that distinguished between the below minimum category and the minimum category was denoted the *minimum cutoff point*, the bookmark that distinguished between the minimum category and the basic category was denoted the *basic cutoff point*, the bookmark that distinguished between the basic category and the proficient category was denoted the *proficient cutoff point*, and the bookmark that distinguished between the proficient category and the advanced category was denoted the *advanced cutoff point*. The relation between the cutoff points and the performance categories is illustrated in Figure 1.

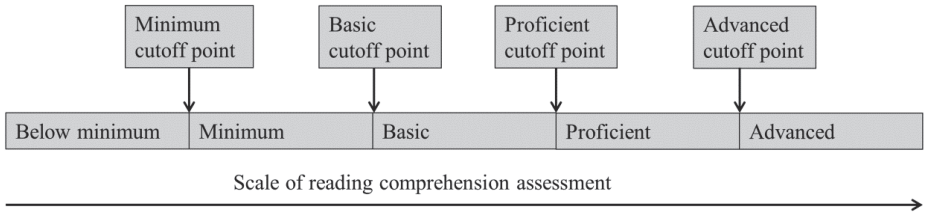


Figure 1. Illustration of the Four Cutoff Points and the Five Performance Categories

During the standard setting meeting, the participants received an introduction to the purpose of the procedure. Next, the cutoff points were made more concrete by presenting relevant descriptors. To facilitate the dialogue about the cutoff points, fictive "example-students" were created, the names of whom alliterated with the cutoff points. Indicators on the quantity of help that these different students received in the classroom were presented. In addition, the panelists were provided with references to actual test performance of the current student population in terms of percentiles. An overview of these different indicators is provided in Table 2.

Table 2

Description of the Four Cutoff Points

Cutoff point	Name	Amount of instruction	Percentile of population at this level
Minimum	Michelle	Extended instruction plus additional remediation	P10
Basic	Benny	Extended instruction	P25
Proficient	Patricia	Regular instruction	P50
Advanced	Arthur	Additional challenging material	P75-P90[13]

During the meeting, the panelists practiced the use of the OIB and standard setting materials[14]. Commonly, items in the OIB are ordered so that each item is more difficult than the previous item. Because reading comprehension assessment items apply to a specific text, the items belonging to the same text were grouped together in our OIB. As a result, the most difficult item of one text could be more difficult than the easiest item of the following text. This aspect of the OIB was specifically attended to. For more information on the training our panelists received, see Deunk, van Kuijk, and Bosker (in press).

After the training in the materials, three rounds were used to derive at the final cutscores. In the first round, panelists studied the OIB individually and set cutscores pertaining to each of the four cutoff points. In the second round, the panelists discussed these cutscores in 17 small groups (consisting of three to five members; colleagues from the same school were not allocated to the same discussion groups). Panelists were asked to share and discuss their cutscores. At the end of this round, panelists could reset the cutscores. We explicitly stated that the groups were allowed to reach consensus with respect to their cutscores but that this was not mandatory. At the end of the second round, we collected the forms on which the cutscores were reset and calculated the median cutscore for each of the four cutoff points. In the third round, the median cutscores for each of the four cutoff points were presented graphically together with performance information of the population under study. This performance information pertained to the actual test results attained by respectively the P10, P25, P50, P75, and P90 of the second and third grade students

[13] Specifically for the advanced level, the panelists could consider students that perform at the 75th or even at the 90th percentile; this level was defined more broadly as it was not pre-defined which percentile was the most accurate representation of students that receive additional challenging materials in class.

[14] Some educational professionals are very familiar with the values on the assessment scales used in the reading comprehension assessment used here. Therefore, the scales used in the first two standard setting rounds on which the cutscores would be identified had been transformed by adding 100 points to the values for reading comprehension. All analyses in this chapter, however, are done using the original scale.

of the Netherlands (these percentiles had been previously discussed with the panelists, see Table 2). At the end of the third round, panelists could reset their cutscores. The medians of these cutscores (set for each of the four cutoff points) were used to determine the final cutoff scores and create the five performance categories. The categories were created as follows: when a student's test score was equal to or higher than the median cutoff score for the minimum cutoff point, performance was labeled to fall into the minimum category. In a similar manner, this labeling procedure was conducted for the basic, proficient, and advanced performance categories. Test scores that fell below the median cutscore for the minimum cutoff point were labeled as below minimum.

3.3.3. The Ordered Item Booklet for reading comprehension

The items from the OIB came from the Cito standardized reading comprehension assessments which have been developed by the Netherlands Institute for Educational Measurement. These standardized assessments form part of the assessment system (LOVS) most widely used in the Netherlands; it has been employed in approximately 85 percent of the Dutch primary schools (Inspectorate of Education, 2010b). In the Cito reading comprehension tests, students are asked to read several texts and answer multiple-choice questions referring to the word, the sentence, and the text levels. These tests are administered from grade one to grade six. Both the validity and reliability of these tests are considered sufficient: their reliability is above 0.89 (Cronbach's alpha) for the grades under study (Feenstra et al., 2010) and the tests have been approved by the Dutch National Committee of Tests and Testing which is responsible for the review of tests (COTAN).

The OIB was constructed in collaboration with the Netherlands Institute for Educational Measurement, and consisted of 30 items that originated from different tests[15]. Two versions of the booklets were made; one for grade two and one for grade three. Both booklets contained items ranging from very easy to very difficult, and the average student's proficiency score lay well within the range of the easiest and the most difficult item in the OIB. For the second grade booklet, the difficulty of the items ranged from -18 to 54 while the average assessment score of a student at the end of grade two is 13.2. For the third grade booklet, the difficulty of the items ranged from -5 to 60 while the average test score of a student at the end of third grade is 30. The negative symbol (-) pertains to the fact that the Cito assessment scale ranges from -87 to +147 (end of grade 1 to mid-term grade 6); the negative symbol should not be interpreted as having any negative connotation.

[15] Each assessment (conducted at the end of first grade, mid-term second grade, end of second grade, mid-term third grade, and etcetera) contains a start test and follow-up test, in which the follow up contains an easy and a difficult version. The items came from all of these tests (i.e., the start test, the easy follow up and the difficult follow up) using assessments conducted at different time points (end of grade one, grade two, etcetera).

3.3.4. Instruments and variables

The following instruments and variables were used.

Evaluation form (procedural validity): At the end of the standard setting procedure, panelists were asked to fill in an evaluation form which contained nine questions pertaining to different aspects of the standard setting procedure (see Appendix 6). In the current chapter, the results of the questions which pertained to the cutscores' procedural validity were used. These questions were formulated as statements containing Likert-scale response options (*yes – a little – no*) and included room for panelists to elaborate their answers. In order to assess the degree of explicitness, participants were asked whether they considered the explanation on how to set cutscores to be clear. In order to assess the practicability, panelists were asked whether they considered the execution of each of the three rounds to be clear. Last, we asked whether the panelists considered their own cutscores to be set in a well-considered manner for each of the three rounds; these answers pertain to the deliberateness with which the participants have set cutscores.

The cutscores for the four different cutoff points across the three rounds (internal validity): The cutoff scores for the second and the third grade were set on special forms. For the investigation of the panelists' adaptations across rounds, difference scores were computed by subtracting the first round cutscores from the second round cutscores, and by subtracting the second round cutscores from the third round cutscores. In order to test the degree of change, absolute difference scores were used in the analyses. To investigate the correspondence between cutscores and empirical performance data, panelists' final cutscores (i.e., set at the end of the third round) were compared to the empirical data (which was presented at the beginning of the third round). To investigate the interpanelists' agreement (i.e., whether or not there was a decrease in variability across rounds), we studied the total variance in cutscores across the three rounds.

3.3.5. Analyses

In order to evaluate the procedural validity of the cutscores, we investigated panelists' feedback with respect to the criteria of explicitness, practicability, and deliberateness with help of descriptive analyses.

In order to assess the interval validity of the cutscores, the results for both the second and third grade were studied at the end of the first, second, and third round of the standard setting procedure. Preliminary inspection of the cutscore distributions showed non-normal distributions and this non-normality of the distributions was confirmed by the Shapiro-Wilk's *W* test of

normality (Sheskin, 2004)[16]. As a result, rather than presenting and comparing descriptive statistics such as the means and standard deviations of the cutscore distributions, we used their non-parametric equivalents; being the median and Interquartile Range (IQR; a measure of dispersion which is calculated by taking the difference between the first quartile and the third quartile). To assess whether the panelists adapted their cutscores across the rounds (the first criterion to evaluate the cutscores' internal validity), it was tested whether the absolute difference scores were statistically different from zero. For this purpose, we conducted the one-sample Wilcoxon signed rank test which is the non-parametric equivalent of the one-sample *t*-test (Sheskin, 2004). In total, 16 tests were conducted; the differences between the first and the second round and between the second and the third round were tested, for each of the four cutscores in each of the two grades. We originally set the significance level at $\alpha = 0.10$ to account for loss in power due to the relative small size of the sample that participated in the current study, and corrected for chance capitalization using a Bonferroni correction which resulted in an alpha level of $\alpha = .10 / 16 = .00625$. The one-sample Wilcoxon tests were conducted one-sidedly as the median of the absolute difference scores could only be similar to or larger than zero.

The correspondence between panelists' cutscores and performance data (the second criterion) was studied by testing whether panelist's ratings for each cutoff point were significantly different from their "empirical equivalent" as presented in the round of performance data – i.e., the actual test score of the P10, P25, P50, P75, and P90-percentiles. For the advanced cutoff point, it was not defined beforehand whether this cutoff point should pertain to the P75 or the P90, hence, the results for this cutscore were compared twice. In total, the one-sample Wilcoxon test was conducted 10 times (five times for the two grades), and the results were compared to a corrected alpha level of $\alpha = .10 / 10 = .01$. In these analyses, the one-sample Wilcoxon tests were conducted two-sidedly as no specific results were hypothesized.

In order to investigate the interpanelists' agreement (the third criterion of cutscores' internal validity), we studied the variance across the three rounds as a decrease in variation would indicate that judgments became more consistent. For this purpose, the total variance in cutscores per round was estimated, while controlling for fixed effect (systematic) differences between the four different cutscores. Here, the unit of analysis is the cutscore. Teachers were asked to set four cutscores, thus there were a maximum number of (39 panelists x 4 =) 156 units (i.e., cutscores) per round for second grade and a maximum number of (28 panelists x 4 =) 112 units per round for third grade. For each round and for each grade, we estimated a separate model in order to identify the total variance in cutscores (the total variance resulting from variation between panelists as well as within panelists as they set multiple cutscores). A confidence interval (CI)

[16] All cutoff score distributions significantly deviated from normality, with the exception of the distribution for the basic cutscores in third grade round 1 and 2, the proficient cutscore in third grade in round 1, and the advanced cutscore in third grade round 3.

was constructed around this total variance-measure. In order to assess whether variation decreased, we compared the total variances in cutscores across the rounds and checked the overlap between CI's.

3.4 Results

First, we evaluate the criteria on the cutscores' procedural validity using the feedback of the participants. Not all panelists filled in an evaluation form and not all questions were answered, hence the *n* fluctuates somewhat per question in Table 3.

Table 3

Summary of the Participants' Responses to the Questions on the Evaluation Form

	Yes		A little		No		Total
	n	(%)	*n*	(%)	*n*	(%)	*n*
Explicitness							
Was the explanation of the standard setting procedure clear to you?	46	(79)	12	(21)			58
Practicability							
Was it clear to you how to carry out round one?	51	(91)	5	(9)			56
Was it clear to you how to carry out round two?	50	(89)	5	(9)	1	(2)	56
Was it clear to you how to carry out round three?	51	(93)	4	(7)			55
Deliberateness							
Do you find your own cutscores from round one to be well-considered?	45	(83)	9	(17)			54
Do you find your own cutscores from round two to be well-considered?	37	(70)	15	(28)	1	(2)	53
Do you find your own cutscores from round three to be well-considered?	38	(72)	13	(24)	2	(4)	53

The vast majority of the participants considered the procedure to be clearly explained, easily implemented, and their own input to be well-considered. The answers to the questions on deliberateness, in comparison to the other questions, received the lowest percentage of confirmatory responses. Three panelists that responded to these questions with either *a little* or *no* provided additional comments on their evaluation forms: one panelist indicated that she felt she was influenced by the opinions of the group members in the small-group discussion round, and two panelists (a principal and an internal support coordinator) indicated that they had limited experience teaching this age group.

Next, we discuss the evaluation of the criteria on the cutscores' internal validity. In Table 4, we report the four median cutscores for each of the three rounds of the standard setting procedure for both grades. In the first round, not all panelists managed to set cutscores for all four cutoff points, hence, the *n* is found to fluctuate. Furthermore, several panelists did not hand in the last form with their final cutscores (they accidentally took it home), which is why the *n* is lower in the third round in comparison to the first two rounds.

Table 4

Summary of the Cutscores set across the Standard Setting Rounds for Second and Third Grade

	Minimum			Basic			Proficient			Advanced		
Rounds	Median	IQR	*n*	Median	IQR	*n*	Median	IQR	*n*	Median	IQR	*n*
						Second grade						
1: Individual round	-7	4	38	1	9	37	20	8	37	25	8	31
2: Discussion round	-7	0	39	1	4	39	14	9	39	30	6	39
3: Empirical data round	-7	2	33	1	4	33	12	8	33	27	7	33
						Third grade						
1: Individual round	5	7	28	19	9	26	30	8	26	43	10	25
2: Discussion round	5	5	28	18	7	28	28.5	3	28	43	12	28
3: Empirical data round	12	8	26	20	3	26	30	5	25	45	8	25

For the second-grade cutscores, the changes in the cutscores for the proficient and advanced cutoff points were substantial when comparing the first to the second round as presented in Table 4. After the small-group discussions, a relatively large decrease in the median of the proficient cutscore is found, while there is an increase in the median of the advanced cutscore. After being provided with performance information in the third round, the median cutscores for both cutoff points were lower than in the previous round. For the minimum and basic cutoff point, the median cutscores remained stable across the rounds. For grade three, the median of the basic and the proficient cutscore decreased slightly from first to second round. Furthermore, there appeared to be a consistent increase in cutscores for all the four cutoff points from the second to the third round following the presentation of performance information; particularly the change in the median of the minimum cutscore was substantial. When comparing the two grades, the impact of the small-group discussions was the largest for the second-grade cutscores given the substantial changes for the "higher" cutoff points, while the impact of the performance information was the largest for the third-grade cutscores.

In Table 4, the median cutscores of several cutoff points did not change across rounds. This does not entail that individual panelists did not revise their ratings. When taking a closer look at the adaptations made by individuals, the largest downwards adjustment was found for the advanced cutscore: at the end of the second round, a panelist had set the cutscore 19 points lower in comparison to the first round cutscore. The largest upwards adjustment was found for the proficient cutscore: at the end of the third round, a panelist had set the cutscore 23 points higher in comparison to the second round. Thus, sizable adaptations in cutscores can be found when looking at the data of individual participants. We analyzed these adaptations using absolute difference scores. In Table 5, an overview of these absolute difference scores is provided.

Table 5

Summary of the Absolute Difference Scores across the Standard Setting Rounds for Second and Third Grade

	Minimum			Basic			Proficient			Advanced		
Difference	Median	IQR	*n*	Median	IQR	*n*	Median	IQR	*n*	Median	IQR	*n*
						Second grade						
Round 2 – 1	1	2.25	38	4	6.5	37	2	7	37	2	8	31
Round 3 – 2	0	2	33	0	5	33	1	8	33	1	4.5	33
						Third grade						
Round 2 – 1	0	1.75	28	0	5	26	1	6	26	1	8.5	25
Round 3 – 2	2.5	7	26	2	5.25	26	1	6	25	2	8.5	25

The aforementioned pattern that the impact of the group discussions appeared larger in second grade and the impact of performance data appeared larger in third grade becomes more apparent from the results presented in Table 5. Interestingly, while the median cutscore for the basic cutoff point in second grade was stable from round one to round two (with a value of 1, see Table 4), from the results in Table 5 it is evident that quite some adaptations have been made for this cutoff point. The median of the absolute difference score is 4. For the median cutscore of the minimum cutoff point, a similar trend is visible although the median of the absolute difference score is somewhat smaller. As the adaptations across rounds are an indicator of internal validity, it was tested whether the 16 medians of the absolute difference scores were each significantly larger than a median of 0 using the one-sample Wilcoxon test. For all 16 tests, the results were $p = .001$ or smaller, and these results are significantly smaller than the corrected $\alpha = .10 / 16 = .00625$.

The second criterion for evaluating the internal validity of cutscores pertained to whether the final cutscores were significantly different from their empirical equivalent as presented in the third round. In Table 6, the medians of the final are presented together with the actual test score of their empirical equivalent. For the advanced cutoff point, it was not predefined whether this cutoff point should pertain to P75 or P90; its results were compared twice. The one-sample Wilcoxon test was conducted 10 times, and the results compared to the corrected $\alpha = .10 / 10 = .01$. The final values (as set during our PD program) that differed significantly from the population values are denoted with an asterisk in Table 6.

Table 6

Comparison between the Final Median Cutscores and the Performance of Different Population Percentiles in Second and Third Grade

	Grade 2		Grade 3	
	PD program	Population	PD program	Population
Minimum (P10)	-7*	-5.9	12	12.5
Basic (P25)	1	2.3	20	20.5
Proficient (P50)	12	13.2	30	30
Advanced (P75)	27*	23.4	45*	39.5
Advanced (P90)	27*	33	45	48.5

* $p < .01$.

From the results presented in Table 6, it can be seen that the minimum, basic, and proficient cutscore were similar to or slightly lower than their empirical equivalent. The median for the

minimum cutscore in the second grade was found to be significantly lower than its empirical equivalent. The median cutscore for the advanced cutoff point was found to be in between the performance demonstrated by the 75th and 90th percentile in the population. For second grade, the median cutscore for the advanced cutoff points differed significantly from both reference points. For third grade, the median cutscore for the advanced cutoff point differed significantly from the 75th percentile's score.

The third criterion for evaluating the internal validity of cutscores pertained to whether or not the variation in cutscores decreased across the rounds as an indication of increasing agreement among the panelists. In Table 7, the results of these analyses are presented. Here, we estimated the total variance in cutscores per round and per grade, while controlling for fixed effect (systematic) differences between the four different cutscores.

Table 7

Comparison of the Total Variance in Cutscores across the Three Rounds per Grade

	Grade 2		Grade 3	
	Variance (*SE*)	95% CI	Variance (*SE*)	95% CI
Round 1	39.63 (4.69)	[30.45; 47.32]	39.11 (5.40)	[28.53; 47.97]
Round 2	22.11 (2.50)	[17.20; 26.21]	30.62 (4.09)	[22.60; 37.33]
Round 3	18.57 (2.29)	[14.09; 22.31]	17.01 (2.38)	[12.34; 20.92]

From the results presented in Table 7, it can be seen that the total variance in cutscores (being the variation between panelists as well as within panelists as they set multiple cutscores) decreased across the different rounds for both grades under study. As the confidence intervals of the first and the third round variances do not overlap, this is a conservative indication – though not a formal test – of a significant difference across the rounds for both grades.

3.5 Conclusion and discussion

In the current chapter, it was investigated to what extent standard setting results met the criteria for the internal and procedural validity of cutscores as we wanted to evaluate the assumption that the cutscores and associated performance categories (defined during our PD program) were valid. By doing so, we aimed to provide an empirical example that would help make the guidelines on standard setting evaluations and validity issues more tangible. The procedural validity of cutscores was investigated with help of panelists' feedback by looking into a) the procedure's explicitness, b) the procedure's practicability, and c) the panelists'

deliberateness. Panelists' responses confirmed these criteria; the questions on an evaluation form targeting these criteria were answered positively by 70 to 93 percent of the panelists. We consider these findings in support of the procedural validity of cutscores.

To evaluate the internal validity of cutscores, we targeted criteria pertaining to d) the panelists' adaptations across rounds, e) the correspondence between cutscores and empirical performance data, and f) the interpanelist agreement. It was found that the adaptations across rounds were significantly larger than zero. According to the standard setting literature, adaptations indicate that panelists refined their judgments with help of the different rounds in the procedure. For this criterion, the internal validity appears to be supported. For the comparison between cutscores and empirical performance data – the second criterion -, it was found that the median cutscore of the minimum cutoff point in second grade was significantly lower than the test score of its empirical equivalent (P10). The basic and proficient median cutscores did not differ significantly from their empirical equivalent (P25 and P50, respectively). For the advanced cutscores in both grades, the cutscores were placed in between P75 and P90. The criterion at hand aims to ensure that standard setting results are realistic; the majority of the results reported here appear to confirm this criterion although the median of the minimum cutoff point in second grade was set relatively low. Yet the interpretation of these results is less straightforward considering the performance improvement context in which cutscores are commonly set (i.e., in our PD program as well as in other countries using standards). It might be preferred that cutscores are set at a level which is slightly higher than their current empirical equivalent. On the other hand, attainment targets are commonly formulated in terms of the performance categories. The performance categories themselves might not need to be ambitious, but perhaps the attainment targets should be. In our program, teachers set attainment targets - i.e., performance goals - for their own students based on these performance categories. These performance goals could be set at a more ambitious level given a student's capabilities. The relation between (ambitious) goals and students' performance will be addressed in the following chapter, but more research is needed to further investigate this issue of realistic versus ambitious cutscores (also in Cizek & Bunch, 2006; Hambleton & Pitoniak, 2006). The third criterion on the cutscores' internal validity, pertaining to the interpanelist agreement, was confirmed by the finding that there was a decrease in cutscore variance across the rounds. Summarizing, the internal validity of cutscores established by our panelists appears to be supported.

In this section, we would like to acknowledge several limitations in this study. For the investigation on procedural validity, not all panelists filled in an evaluation form – our results might thus only pertain to a specific subsample of panelists. Moreover, the way in which the questions were framed might be considered directive and they have been formulated in such a way that they might have elicited socially desirable responses (see Cozby, 2003). It therefore remains unclear whether the panelists truly considered the standard setting procedure to be well-

implemented. This could have been dealt with by asking open-ended questions or to word the questions in such a way that consistent agreement is unlikely. Another aspect we would like to address pertains to the cutoff point for the advanced category. This category was defined more broadly as it was not pre-defined whether "receiving additional challenging material in class" (presented as a descriptor in Table 2) should pertain to students at P75 or at P90. The finding that the median cutscore is precisely in between P75 and P90 is possibly a direct consequence of our own general approach to this category; more precise directions (e.g., by clearly selecting either the P75-student of the P90-student for this category, as we did for the other categories) might have yielded different results. Furthermore, an important note previously addressed is that the evidence on validity is conditional. As there are no ''true" cutscores, it is impossible to evaluate whether the cutscores have been set at the accurate level (also in, for example, Pant et al., 2009). In addition, what makes the evaluation of standard setting procedures more complex is that there are no *absolute* criteria that can be employed. Standard setting is based on human judgment (Berk, 1986), and so is the evaluation of standard setting results. For instance, how would we judge the results if only 60 or even 50 percent of the panelists would indicate that they felt their cutscores were set at a well-considered level (our criterion of deliberateness)? Would this still be considered a sufficient support of procedural validity? Concrete descriptions of evaluation results that are considered to be either valid or invalid would be considered a valuable contribution to help standard setting evaluators.

Given the complex and judgmental nature of the standard setting procedure, it is important to evaluate the validity of cutscores each time a standard setting procedure is used. Our study can be considered an illustration on how such an evaluation can be conducted, but it simultaneously illustrates that this field is in need of further work.

4. Teacher-set performance goals and relations to student achievement

Abstract: As part of a teacher professional development program, the participating teachers were asked to set a goal for each of their students pertaining to these students' reading comprehension performance at the end of the school year. In order to assist teachers in the goal setting task, a multistep procedure (which incorporated performance data analysis and team discussion) was developed to help teachers reflect on and reconsider the goals' appropriateness before deciding on the final goal. In the current chapter, we assessed the use of this procedure by evaluating change across the procedure, i.e., whether the final goals were equal or different to the goals the teachers had set at the beginning of this procedure. In addition, we evaluated the relation between the final goals and these students' achievement by focusing on a) the attainment of the goals, and b) whether the goals were significant predictors of student achievement while controlling for relevant student and classroom level covariates. In the analyses on the use of the multistep procedure, a significant amount of change across the procedure was found, which was considered to be indicative of the final goals' deliberateness. Furthermore, 79 percent of students had attained their goal by performing at the desired level or higher. Moreover, the performance goals were found to be significant predictors of performance, and higher goals were associated with higher results. The positive effect of high goals on achievement was even stronger for initially low-achieving students.

4.1 Introduction

Currently, there are concerns on the early reading proficiency of Dutch students that call for attention and action (Houtveen & Van de Grift, 2012; Inspectorate of Education, 2007; Van Berkel et al., 2007). Insufficient results of Dutch students on both international and national reading assessments have been attributed to the fact that, to schools and teachers, it was unclear what students should know and do at certain time points (Council of Education, 2007; Inspectorate of Education, 2011; Ministry of Education, 2010). Aiming to improve the reading performance of Dutch students in second and third grade (student age: approximately 7 to 9 years old), we developed a teacher Professional Development (PD) program in which goals played an important role. As part of this PD program, teachers were asked to set a performance goal for each of their students. This goal setting task was assumed to improve teacher instruction, as setting goals helps to focus the attention toward (the attainment of) desired results. Subsequently, this improved instruction was assumed to result in improved student achievement. The hypothesized positive relation between goals and achievement is based on findings from goal setting theory. Studies in this field have identified such relations, particularly in situations in which goals are set at an ambitious level (Locke & Latham, 1990; 2002). Similar results are reported in the school effectiveness literature (Scheerens & Bosker, 1997) and the teacher expectancy literature (Jussim & Harber, 2005; Rosenthal & Jacobson, 1968; Rosenthal, 1987) in which ambitious achievement expectations are associated with higher student results. This has found to be particularly the case for initially low-achieving students (Good & Brophy, 2003). Yet the level of ambition should not be taken to an extreme when setting goals: the most motivating goals are those that are difficult but not too difficult. Erez and Zidon (1984, in Locke & Latham, 1990) found that performance leveled off or decreased when limits of ability were reached, or when the commitment to a difficult goal lapsed.

The goals in our PD program were formulated by the teachers. For each of their students, they selected one of five *performance categories*. These performance categories had been defined by the participating teachers in an earlier stage of the PD program with help of a *standard setting procedure*. A specific feature of these performance categories was that they pertained to test score intervals on the scale of the end-of-the-school year standardized reading comprehension assessment. The advantage of setting goals in terms of categories which had been linked to an assessment ('I want Billy to attain a score within the *proficient* category and Julie to attain a score within the *advanced* category on the standardized reading comprehension assessment which is conducted at the end of the school year') was that the attainment of these goals would easily be established by conducting the assessment in class. In order to assist teachers in their goal setting task, we developed a *multistep procedure* which incorporated performance data analysis and team discussion to help teachers reflect on and reconsider the goals' appropriateness before deciding

on its final version (following recommendations of the data use literature: e.g., Schildkamp & Kuiper, 2010).

The teacher-set performance goals have played a key role throughout the PD program and are the focus of the current chapter. First of all, we are interested in the use of the multistep procedure to get an indication of whether the teachers have set their goals in a well-considered way. Next, we investigate the relation between goals and students' results using two approaches, namely by focusing on the extent to which teachers have attained their own goals and by investigating whether the goals are associated to students' growth in reading comprehension.

In the paragraphs below, we will elaborate on the rationale behind working with goals and the relation between high goals and high achievement. Related findings from the field of teacher expectancy research are discussed as well. Subsequently, information is provided on how the student-specific performance goals have been set; teachers first participated in a standard setting procedure to create the performance categories, and then participated in the multistep procedure in which, at the end of the procedure, they were asked to set a performance goal for each individual student. The details of the current study and the research questions will be provided before continuing to the methods section.

4.2 Theoretical framework

4.2.1. Working with goals

Working with goals has generally been proven to be effective for enhancing performance. Setting goals leads to a clearer notion of how desired outcomes can be attained, and it directs the focus toward the attainment of these desired outcomes (Fuchs et al., 1985; Fuchs et al., 1989; Locke & Latham, 1990; 2002). For example, in the study of Fuchs, Fuchs and Deno (1985), teachers were asked to set goals for their students. In post-study interviews, the participating teachers indicated that students' development could more accurately be monitored due to the focus on whether students were making sufficient progress toward attainment of the goal.

When working with performance goals, these goals should be defined at a level that challenges teachers and their students. More ambitious goals are associated with higher student performance (e.g., Fuchs et al., 1985) because these ambitious goals lead to greater effort and persistence, and they direct the attention toward goal-related activities (Locke & Latham, 2002). The mean effect sizes in the meta-analysis of Locke and Latham (1990), containing studies from both organizational and educational settings, ranged from $d = .52$ to $d = .82$ when comparing the effects of difficult to easy-to-reach goals. This importance of ambitious goals applies to all the students in a teacher's class. Weaker performing students are frequently presented with less challenging goals and tasks in order to avoid frustration (Good & Brophy, 2003). Yet this conduct has been proven to negatively affect students' performance as these students are

provided with less opportunity to learn. Furthermore, it has been demonstrated that this conduct negatively influences students' self-confidence, as these students are aware of the fact that they are receiving less demanding tasks in comparison to classmates (Houtveen, Mijs, Vernooy, & Roelofs, 2000; Rubie Davies, Hattie, & Hamilton, 2006). Such students benefit from high goals especially (Good & Brophy, 2003).

The association between ambitious demands and higher student results has also been discussed in the teacher expectancy literature (e.g., Harris & Rosenthal, 1985; Jussim, Eccles, & Madon, 1996; Madon, Jussim, & Eccles, 1997; Rubie Davies et al., 2006). Induced expectation experiments, in which teachers were provided with manipulated information on their students' potential, have demonstrated that students whose teachers have been led to hold high expectations achieved more than other students. The 'Pygmalion in the classroom' study of Rosenthal and Jacobsen (1968) is the most well-known study of this so-called *self-fulfilling prophecy effect*. But also studies on naturally formed expectations of teachers show the same trend (de Boer, Bosker, & van der Werf, 2010; Madon et al., 1997; McKown & Weinstein, 2008; Rubie Davies et al., 2006; Rubie Davies, 2006; Van der Hoeven-van Doornum, Voeten, & Jungbluth, 1989). Some refinement of the research results on teacher expectations is in place though. Jussim and Harber (2005) have reviewed the studies in this area and have concluded that self-fulfilling prophecy effects are (although occasionally large) typically small, with $r = .1$ or $r = .2$. The relationship between teacher expectations and student achievement is frequently found to be larger but this is because teachers are accurate in their expectations of students; "predicting, but not causing student achievement" (p. 138). While *prediction* stems from accurate expectations, self-fulfilling prophecies result from inaccurate expectations *causing* certain results to be attained and thus becoming accurate. This is a crucial difference.

The literature on teacher expectancies and self-fulfilling prophecies has been discussed here as we consider teacher-set goals to be a combination of a) teachers' expectations, and b) teachers' ambitions. These goals are formulated by combining "what the individual thinks *can be* achieved and what he or she *would like to* achieve or thinks *should be* achieved" (Locke & Latham, 1990, p. 122). Goals should be set at an ambitious level but if goals have been set too high, teachers and their students will be unsuccessful in attaining this goal. Such failure can negatively affect students' self-confidence (Seifert, 2004) or teachers' self-efficacy beliefs (Skaalvik & Skaalvik, 2007).

4.2.2. A teacher PD program targeting goals based on performance categories

During the teacher PD program, we asked teachers to set goals that are *difficult but not too difficult* - referring to the prior reported results of Erez and Zidon (1984, in Locke & Latham, 1990) - in which we acknowledged that this was a complex task. There is no known optimum between expectation and ambition for each student to which the teacher-set goal can be compared.

However, we aimed to assist teachers in setting an appropriate goal by investing considerable time and effort in the goals' development and by stimulating the use of different sources of information while the teachers were setting their goals.

The entire PD program contained a total of nine after-school meetings with accompanying homework assignments. Second- and third-grade teachers participated in the PD program, as well as their school's principal and internal support coordinator. All participants took part in the standard setting procedure which was conducted during a plenary meeting in November 2011. It was in this particular meeting that the performance categories (i.e., the test score intervals on which the goals were based) were defined. The meeting in which the student-specific performance goals were set (i.e., at the end of the multistep procedure) was conducted at the individual schools in November and December of 2011. The goals which were set during this goal setting meeting pertained to test results on the reading comprehension assessment which would be conducted in June/July of 2012. Throughout the remainder of the PD program, teachers received training in the use of the student monitoring system (to track students' progress) and in relevant instructional skills and knowledge in reading comprehension in order to facilitate the attainment of the performance goals.

In the following paragraphs, the standard setting procedure and the multistep procedure will be described. Studies on the interpretation of test scores by teachers, albeit limited in their number, might warrant against the use of test scores for goal setting purposes due to their frequent misinterpretation (see e.g., van der Kleij & Eggen, 2013). Yet as the teachers in our PD program studied concrete reading comprehension items as part of the standard setting procedure, and were trained in understanding the link between items and test scores, this was not considered to be problematic.

4.2.2.1 Defining performance categories using a standard setting procedure

Defining the performance categories was a preliminary step before the teachers set goals for their own students' performance. The term *standard setting* is used to refer to the procedure in which performance categories are created, by setting cutoff points that define the boundaries of these categories (Hambleton & Pitoniak, 2006). We used the Bookmark procedure (Karantonis & Sireci, 2006; Mitzel et al., 2001) in which the participants were asked to place bookmarks at the appropriate cutoff point between consecutive categories in order to create multiple performance categories (associated to different levels of proficiency). This task of identifying suitable cutoff points was facilitated through the use of the *Ordered Item Booklet* (OIB) in which items from standardized reading comprehension assessments were ordered in such a way that they increased in their difficulty. In our program, the participants were asked to place four bookmarks in order to create five performance categories. Each bookmark referred to a certain cutoff point between two

successive categories. The relation between the cutoff points and the performance categories is illustrated in Figure 1.

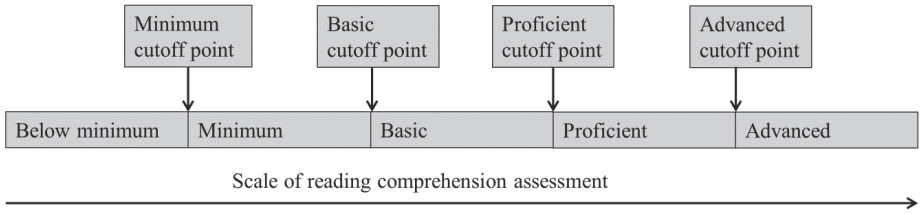


Figure 1. Illustration of the Four Cutoff Points and the Five Performance Categories

At the beginning of the standard setting meeting, an introduction was provided on the purpose of the meeting and the participants received training in the use of the standard setting materials. The four cutoff points were made more concrete by referring to fictive "example-students" that had different levels of reading proficiency and by presenting relevant descriptors pertaining to a) the quantity of help that these students received in the classroom, and b) references to actual test performance of the current student population in terms of percentiles. An overview of the different indicators is provided in Table 1. For more information on the different cutoff points and other aspects of the standard setting training, one is referred to Chapter 3.

Table 1

Description of the Four Cutoff Points

Cutoff point	Name	Amount of instruction	Percentile of the population at this level
Minimum	Michelle	Extended instruction plus additional remediation	P10
Basic	Benny	Extended instruction	P25
Proficient	Patricia	Regular instruction	P50
Advanced	Arthur	Additional challenging material	P75-P90[17]

[17] Specifically for the advanced level, the panelists could consider students that perform at the 75th or even at the 90th percentile; this level was defined more broadly as it was not pre-defined which percentile was the most accurate representation of students that receive additional challenging materials in class.

During the standard setting meeting, the participants studied items from the Cito standardized reading comprehension assessments (developed by the Netherlands Institute for Educational Measurement). As the items corresponded to test scores on the end-of-the-year Cito reading comprehension assessment, these cutoff points could thus be established in terms of test scores. In the standard setting literature, cutoff points are also referred to as *cutscores* or *cutoff scores*. After three rounds of standard setting, being 1) individual work, 2) group discussion, and 3) investigation of students' actual performance, the final cutscores were determined for both second- and third-grade reading comprehension.

As the cutscores identify the boundaries of the performance categories[18], one can establish the test score intervals which are associated to these categories. The performance categories and their associated test score intervals are presented in Table 2. Certain intervals contain scores with a negative symbol (-). This is because the scale of the Cito reading comprehension assessments from end of grade 1 up to mid-term grade 6 ranges from -87 to +147. This negative symbol should not be interpreted as having a negative connotation; -87 is simply the (arbitrary) starting point of this scale. In the general population of students who take the Cito reading comprehension assessment, the average proficiency score of a student at the end of grade two is 13.2 and the average proficiency score of a student at the end of third grade is 30.

Table 2

The Performance Categories and their Associated Range of Test Scores on the Cito Reading Comprehension Assessment

	Test score intervals	
Performance category	Second grade	Third grade
Below minimum	≤ -8	≤ 11
Minimum	-7 – 0	12 – 19
Basic	1 – 11	20 – 29
Proficient	12 – 26	30 – 44
Advanced	≥ 27	≥ 45

After creating the performance categories, teachers were asked to set performance goals for their students by selecting the most appropriate performance category for each student in their

[18] The performance categories were created as follows: when a student's test score was the same or higher than the median cutscore for the minimum cutoff point, performance was labeled to fall in the minimum category. In a similar manner, this labeling procedure was conducted for the basic, proficient and advanced performance categories. Test scores that fell below the median cutscore for the minimum cutoff point were labeled as below minimum.

class. To facilitate the setting of a well-considered goal for each student, it was considered desirable that teachers reflected on their own expectations and ambitions with help of different sources of information. For this purpose, we developed the multistep procedure, which we will now discuss. In the section below, the name of each step in this procedure is presented in italics.

4.2.2.2 The multistep procedure for setting performance goals

The first step of the procedure was part of a homework assignment. The participating teachers completed this assignment prior to the meeting in which the performance goals were set. In this first step, they were asked to predict the end-of-the-year assessment result by selecting one of the five performance categories for each of their students (e.g., 'If I think of my student Julie, I consider her to be a relatively average performing reader. At the end of the school year, I expect her to attain a test score within the proficient category.'). For the remainder of this article, this step is referred to as the *initial goal*.

In the second step of this procedure, also part of the homework assignment, teachers were asked to focus on *students' performance data*. Teachers were asked to consider the student's results on the previous Cito standardized reading comprehension assessments. As it has been identified that teacher expectations (and related, teacher bias) are formed based on a large number of individual student characteristics – including a student's sex, social class, diagnostic labels, and the relationship between the teacher and the student's other siblings (Rubie Davies et al., 2006) -, explicitly focusing teacher's attention on prior attainment was expected to improve the accuracy of teacher's expectations (also noted by Good and Brophy, 2003). A student's performance data would frequently concur with the teacher's initial goal, but there could be a discrepancy between this initial goal and the data (e.g., 'Julie attained an excellent mark on the last standardized reading comprehension test. Perhaps the advanced performance category could be more appropriate for her'). For those cases that the initial goals and the performance data were not aligned, teachers were asked to write down a possible explanation for this inconsistency.

In the third step of the procedure and the last part of the homework assignment, teachers were asked to look up the end-of-the-school year assessment *results of students that were in the current grade in the previous school year*. The results of students they had taught in the previous year could be classified into one of the five performance categories (by allocating the test scores to the correct test score intervals). This was expected to help teachers get a better "feeling" for the differences between the performance categories (e.g., 'If I think of Amy whom I taught last year, she was a student who I consider to be very similar to Julie. On the reading comprehension test which was conducted at the end of the school year, Amy attained a test score that would be classified as advanced').

During the meeting in which teachers set their goals, teachers were asked to discuss the homework assignment with their colleagues, and especially elaborate on the students whose

performance data did not align with teachers' initial goals. This *team discussion* was the fourth step of the procedure. Teacher collaboration and team discussion are considered important steps when analyzing student performance, as teachers can ask each other for help and give each other advice (Lai & McNaughton, 2013; Schildkamp & Kuiper, 2010; Schildkamp, Lai, & Earl, 2013). The participants - being the school's principal and internal support coordinator and the other (second- or third-grade) teachers from their school– worked together in trying to think of appropriate goals and explain possible discrepancies (e.g. a colleague saying: 'I remember teaching Julie last year, and I found her to be a very skillful reader. She is a very quiet student though, who does not draw much attention to herself. Could this be the reason that you initially set a somewhat lower goal for her?').

At the end of the procedure, teachers were asked to set their final performance goals by allocating each student to a performance category (e.g., 'At the end of the school year, I want Julie to perform at the advanced level. I will select the advanced performance category as my performance goal for Julie'). The goals were not communicated to students as their content was considered too abstract for this age group. More details on the different steps in this procedure as well as the relevant PD program's meetings are provided in Appendix 1 of this dissertation.

4.2.3. The current study

In the current chapter, we focus on evaluating the performance goals which have played a key role in the teacher PD program. First we investigate the use of the multistep procedure by focusing on the degree of change across the procedure. If the teachers have taken in and synthesized the information they were provided with across the different steps, these student-specific performance goals should show some variability. Staying with one's initial goal throughout the procedure would call into question the effectiveness of the different rounds aiming to help the teachers to refine their goals. Even though teachers might consider the information acquired throughout the procedure to confirm their initial ideas (thus not changing from initial goal to final goal), small variation in goals across the procedure is considered as a sign that the information has been used whereas static goals do not provide such information. A similar line of reasoning has been applied in the evaluation of the standard setting procedure, as discussed in Chapter 3. For the evaluation of the multistep procedure, we assess the difference between the initial goal and the final goal rather than focusing on change following each separate step; the procedure aims to help teachers set a well-considered goal by incorporating different sources of information. At which exact step a change takes place is not of interest here. As discussed in Chapter 3, the evidence we collect for the evaluation of this procedure is conditional: evidence of change does not necessarily imply that the most accurate goals are set, whereas lack of change may point to less well-considered and thus, probably, less accurate goals (c.f., Kane, 1994). Here,

change across the procedure is considered to be indicative of teachers' deliberateness during the goal setting process.

After focusing on teachers' deliberateness for the goals they have set, we focus on the relation between the teacher-set goals and students' performance. Goal attainment is considered a relevant outcome measure as it was a focal point during the PD program. Yet it might be an inaccurate indication of satisfactory academic growth. A relatively low goal might be attained without challenging the student while a very high goal might stimulate performance even when the goals itself is not reached (a scenario also proposed in the goal setting work of Fuchs, Fuchs, and Deno, 1985). Therefore, we are also interested in whether the goals are significant predictors of student achievement while controlling for relevant student and classroom level characteristics. By doing so, it can be investigated whether there is a relation between the performance goal and student's test results without requiring the attained test score to fall into a certain range of test scores. By using the goals as a predictor in regression analysis while we account for covariates such as prior reading achievement, it can simultaneously be assessed whether higher goals are associated with higher performance. Furthermore, it will be investigated whether the relation between the goals and achievement depends on initial achievement, as particularly the weak achieving students are known to benefit from high goals. The following research questions are addressed in this chapter:

1) To what extent do the teachers' final performance goals differ from their initial goals?

2) To what extent have the teachers attained their final performance goals?

3) To what extent do the teacher-set performance goals predict academic performance, and are higher goals associated with higher performance while controlling for relevant covariates?

4) Does the relation between the performance goal and student achievement depend on students' initial performance?

4.3 Method

To investigate the research questions at hand, pretest posttest designs were used in this study.

4.3.1. Participants

A total number of 19 schools and 33 teachers from the northern part of the Netherlands participated in our PD program. Schools and teachers participated in this study on a voluntary basis: no financial or other compensation was provided. For the current investigation, we only included the teachers who were included in the effect study of the program (discussed in Chapter 2 of this dissertation) and had set performance goals for their own students during the goal setting

meeting. This resulted in a sample of 27 teachers[19] who taught 358 students. The average number of years of experience is 13.9 (*SD* = 11.5). Two of these 27 teachers were male.

The sample of students for whom the teachers had set goals (n = 358) contained the following characteristics: 194 students were second-graders (54 percent), and 164 students (46 percent) were third-graders. Of the 358 students, 166 students were girls (46 percent). Four students (1 percent) had an official indication of Special Educational Needs. Students with lower educated parents are identified as "potentially at risk" in the Dutch educational system[20]. In our study, 28 students (8 percent) had such an indication.

4.3.2. Instruments and variables

The following instruments and variables were used to answer the different research questions in this study.

Teacher-set performance goals: For each student in class, an initial and a final performance goal were set by the teacher during the multistep procedure. The results for these two goals were each placed on an ordinal scale ranging from 1 (*below minimum*) to 5 (*advanced*). For the first research question, the initial performance goal and final performance goals were compared in order to assess change across the multistep procedure. For our statistical analyses on the use of the procedure, we were interested in the percentage of students in a class for whom the final goal was different from the initial goal[21]. This information, referred to as the teacher's goal adaptation average, was used in the analyses of the first research question. The final performance goal–variable (the ordinal variable, ranging 1 to 5) was used in the analyses of the second, third and fourth research question.

[19] In total, 29 teachers were included in the analyses as discussed in Chapter 2. Here, we elaborate on the two cases that were not included in the analysis on performance goals (this chapter). One teacher did not attend the goal setting meeting and set his goals for student performance much later in the school year. Sufficient time for this aspect of the program could thus not be guaranteed, and his class was therefore excluded from this analysis. One teacher set her goals during the original goal setting meeting, but she retired in spring of 2012. She had been a part-time teacher and her *partner-colleague* had attended several of the meetings as well. As we considered this partner-colleague sufficiently trained by the program, we included this group of students in the analyses of the effect of the program. Yet as the performance goals had been set without the involvement of this partner-colleague, we considered that sufficient time and focus for this aspect of the program could not be guaranteed, and therefore excluded this class from the analysis.

[20] For this variable, we made use of the Dutch pupil weight system, in which the weights 0.00, 0.30 and 1.20 have been distinguished based on the parents' education. The number of students with a 0.30 weight and a 1.20 weight were grouped together.

[21] For these analyses, there was no interest in the sign or size of the difference between the two goals (i.e., whether the initial goal was set higher or lower than the final goal resulting in a positive or negative difference, and whether this difference between the initial goal was one category or more) and we thus worked with the count data. A difference between the initial goal and final goal was counted as '1' (regardless of size and direction) and when these were set at the same category, this was counted as '0'. For each teacher, we calculated a goal adaptation average (the sum score of this count data divided by number of students in the class size). General trends in adaptations with regard to sign or size of the difference between the two goals will be described in the results section.

Reading comprehension assessment results: The Cito standardized reading comprehension assessments – of which a selection of items were used in the standard setting's OIB - were used to measure students' reading comprehension skills. Both the validity and reliability of these tests have been considered sufficient: their reliability is above 0.89 (Cronbach's alpha) for the grades under study (Feenstra, Kleintjes, Kamphuis, & Krom, 2010). Their use has been approved by the Dutch National Committee of Tests and Testing, responsible for the review of tests (COTAN). The reading comprehension tests are part of the Cito assessment system (LOVS) which is used throughout the elementary school period. For most subject areas in the Cito LOVS standardized assessment system, assessments are conducted in January and in June. However, the national reading comprehension assessments have a slightly different timing: they are conducted in June in the first grade, in both January and June in the second grade, and only in January from third grade onward. To obtain an 'end-of-the-school year' result for reading comprehension in the third grade, we used the reading comprehension items of an additional multisubject standardized test (also developed by the Netherlands Institute for Educational Measurement), which is conducted in June. Its results are registered on the same scale as those of the regular reading comprehension assessments.

As aforementioned, the assessment scale of the reading comprehension test ranges from -87 to +147 (end of grade 1 to mid-term grade 6). The second- and third-grade assessment results in June 2012 could be classified into one of the five performance categories, as the categories pertained to test score intervals on the assessment's scale. This is how we created the attained performance category-variable, ranging from 1 (*below minimum*) to 5 (*advanced*), to which the final performance goals were compared in our aim to answer the second research question.

For the third research question, the relation between the final performance goal and students' performance on the standardized reading comprehension assessment was investigated with help of regression analyses. Here, the second- and third-grade assessment results in June 2012 were used as reading comprehension posttest data which were predicted using the final performance goals. The results of the Cito standardized reading comprehension assessment of June 2011 were used as pretest data which were controlled for in the analyses (this variable was grand-mean centered to facilitate its interpretation). For our fourth research question (focusing on whether the relation between the goal and achievement depended on students' initial achievement), this pretest data was of particular interest.

Mathematics assessment results: Mathematics performance was controlled for in the prediction of students' reading results (third and fourth research question) in order to incorporate a proxy for general academic ability. The Cito standardized mathematics assessments, which are also part of the Cito LOVS assessment system, were used here. These mathematics tests have been approved by the Dutch National Committee on Tests and Testing (COTAN) as well. Both the tests' validity and reliability are considered sufficient (Janssen et al., 2010): the tests'

reliability is above 0.91 (Cronbach's alpha). The scale of the mathematics assessments ranges from 0 to 169. The June 2011 assessment results were used as a covariate in our analyses (grand-mean centered to facilitate its interpretation). These data were collected at the same time point as the pretest data for reading comprehension.

Other variables used in the analyses of the third and fourth research question were:

Sex: Boys were the reference group for this dummy-coded variable.

Grade: For this dummy-coded variable, second grade was the reference group.

Indication of Special Educational Needs: For this dummy-coded variable, the students without an official indication of Special Educational Needs were the reference category.

Educational level of the parents: Students "not being potentially at risk" were the reference group for this dummy-coded variable.

Multi-grade classroom: This variable was a dummy-coded variable for which a single-grade class was the reference category. This classroom characteristic was taken into account in the regression analyses (as part of our third and fourth research question). The teachers that participated in our program and who taught in multi-grade classes were found to select the high performance categories more frequently than teachers in single-grade classrooms: relatively more students in the multi-grade classes received a proficient or advanced performance goal. This phenomenon was controlled for by incorporating the students' classroom type in the analyses. One classroom was a single-grade classroom for the first half of the school year, and a multi-grade classroom for the second half of the school year. We treated this classroom as a single-grade class as performance goals were set in the single-grade situation.

4.3.3. Analyses

For the first research question (comparing the initial goal prediction and final performance goals), it was tested whether the teachers' goal adaptation average per teacher was significantly larger than zero with help of a one-sample *t*-test (Sheskin, 2004). The one-sample *t*-test was conducted one-sidedly ($\alpha = .05$), as the mean of the absolute difference scores could only be similar to or larger than zero.

For the second research question, we investigated the attainment of teachers' goals using descriptive analyses (comparing the final performance goals with the attained performance categories).

For the third research question, a multilevel regression analysis was performed with the help of the software *MLwiN* (Rasbash et al., 2011), with students (level 1) nested in classes (level 2). It was analysed whether performance on the posttest could significantly be predicted by the performance goal while controlling for the aforementioned covariates (at the student and

classroom level). By doing so, we could study whether a student for whom the teacher had set a higher goal performed better at the end of the school year than the student for whom the teacher had set a lower goal, while taking the other predictors of student achievement into account. The initial goal, step one in the multistep procedure, was deliberately not included in these analyses as we were interested in the effect of the final (i.e., the most deliberate teacher-set) goal.

For the last research question, we added an interaction effect between the pretest and the performance goal to see whether the effect of a (high) goal was stronger for students whose prior performance in reading comprehension was relatively low. For the last two research questions, we hypothesized positive effects of higher goals, hence the significance of these explanatory variables was tested one-sidedly ($\alpha = .05$).

4.4 Results

4.4.1. Results of the comparison between the initial and final performance goal

For the evaluation of the deliberateness of the performance goals, we compared the initial goal to the final performance goal. Several teachers had not completed the homework assignment in which they were asked to select an initial goal. In Table 3, the results are presented for those students whose teachers managed to set both the initial goal and the final performance goal (n = 285).

Table 3

Frequencies of the Performance Categories for the Initial Goal and the Final Goal

	Initial goal		Final performance goal	
Performance category	*n*	(%)	*n*	(%)
Below minimum	10	(3)	6	(2)
Minimum	52	(18)	43	(15)
Basic	99	(35)	81	(28)
Proficient	82	(29)	105	(37)
Advanced	42	(15)	50	(18)
Total	285		285	

The higher performance categories (i.e., proficient and advanced) were selected more frequently at the end of the multistep procedure than at the beginning of this procedure, as can be

seen from the results presented in Table 3. When taking a closer look at the adaptations made for individual students, we see that for more than half of the students (n = 181, being 64 percent) the initial goal and the final goal were set at the same level. For 94 students (33 percent) the difference between initial goal and final goal was one category. For eight students – their goals being set by four teachers -, a difference of two categories was found. Two students received a final goal which differed in three categories in comparison to the initial goal (these students had a different teacher). In the cases that the initial goals differed from the final goals, the final goals were more often higher (for n = 79 students) than lower (for n = 25) in comparison to these initial goals. When taking a closer look at the adaptations made by individual teachers, one teacher did not make any adjustments: her final goals were set at the exact same level as her initial prediction for the three students in her second grade class (this teacher's school was situated in a very rural and sparsely populated area). The maximum number of adjustments was made by a teacher who revised her goals for 17 students in her class of 23. Teachers' goal adaptation average (i.e., the percentage of students in a class for whom the final goal was different from the initial goal) was M = .35, with SD = .19 (Min. = 0, and Max. = .74). This goal adaptation average was significantly larger than zero, tested using a one-sample t-test (t[20] = 8.24, p = .000), indicating significant change across the multistep procedure.

4.4.2. The attainment of the performance goals

The attainment of the performance goals was investigated using the data of 27 teachers and their 358 students. In Table 4, a cross tabulation is presented in which the final goals are compared to the performance categories which have been attained on the posttest.

Table 4

Overview of the Teacher-Set Performance Goals Set and the Attained Performance Categories

	Performance categories attained						
Final	Below minimum	Minimum	Basic	Proficient	Advanced	Total	
Performance goal	*n*	*n*	*n*	*n*	*n*	*n*	(%)
Below minimum	4	0	1	2	0	7	(2)
Minimum	9	4	18	13	4	48	(13)
Basic	5	15	25	38	16	99	(28)
Proficient	0	6	16	70	45	137	(38)
Advanced	0	2	4	18	43	67	(19)
Total (%)	18 (5)	27 (8)	64 (18)	141 (39)	108 (30)	358	

From the results presented in Table 4, it can be seen that the performance goals have been attained for 146 students (presented on the diagonal). This is 41 percent of the student population. It was found that 137 students (38 percent; the numbers presented above the diagonal) attained a higher test score than was expected when considering the goal their teacher had set. For this group of students, the goals had been attained as well. For 75 students (21 percent of students, the numbers presented below the diagonal), test results were lower than expected when considering the goal their teacher had set for them.

4.4.3. Using the performance goals as predictors of student achievement

The relation between the performance goals and students' results was also investigated without requiring the attained test score to fall into a certain range of test scores. Here, we discuss the results of the regression analyses in which the performance goals were used as predictors of achievement. Preliminary data inspection of the reading comprehension assessment results showed that - on average - students developed their reading proficiency more strongly in second grade than in third grade. This trend is also evident in the general population (Feenstra, Krom, & van Berkel, 2007a; 2007b). In Table 5, the descriptives for pre- and posttest are presented per grade. The June 2011 mathematics results, used as a covariate in the analyses, are presented in this table as well.

Table 5

Summary of Students' Test Results on Mathematics Assessment, Pretest, and Posttest for Second and Third Grade

		Math results	Pretest	Posttest
Grade	*n*	*M (SD)*	*M (SD)*	*M (SD)*
2	194	47.16 (14.54)	3.57 (16.16)	23.09 (16.09)
3	164	64.13 (15.90)	21.37 (15.28)	33.39 (15.96)
Total	358	54.93 (17.36)	11.73 (18.07)	27.81 (16.81)

Already in an early stage of the analyses, we found that students' sex, parental education, and indication of Special Educational Need were non-significant predictors of performance after incorporating prior achievement. As we preferred the use of parsimonious statistical models, we excluded these variables from further analyses.

In Table 6, the results of the multilevel analysis are reported. The first model (the start model) contained the intercept and the covariates at student and classroom level but excluded the final

performance goal. In the second model (the main effect model), this latter variable was included. In this way, it could be analyzed whether or not the performance goal added value in the prediction of students' reading results. In the third model, we included an interaction effect between goal and pretest (the interest of our fourth research question). The models presented in Table 6 contain unstandardized coefficients.

Table 6

Multilevel Models Predicting Achievement in Reading Comprehension

	Models					
	Start Model		Main effect model		Interaction Model	
Predictors	Coeff.	*SE*	Coeff.	*SE*	Coeff.	*SE*
Fixed Part						
Constant	27.80*	1.33	19.93*	3.20	22.22*	3.32
Grade 3	-4.01*	1.53	-2.31	1.62	-2.90	1.60
Math performance	0.17*	0.04	0.14*	0.04	0.15*	0.04
Pretest	0.64*	0.04	0.57*	0.05	0.83*	0.13
Multi-grade classroom	3.22*	1.40	3.16*	1.36	3.12*	1.30
Performance goal			1.99[a]	0.74	1.63[a]	0.75
Performance goal x pretest					-0.07[a]	0.03
Random Part						
Variance at classroom level	3.70	3.25	3.25	3.07	2.42	2.79
Variance at student level	107.94	8.35	106.06	8.02	105.20	8.13
Deviance	2701.75		2694.61		2689.75	
No. of teachers	27		27		27	
No. of students	358		358		358	

* $p < .05$, two-sided.
[a] $p < .05$, one-sided.

When comparing the start model to the main effect model, one can see that the performance goal is a significant predictor of achievement, and higher performance goals are indeed

associated with higher performance as the sign of the regression coefficient is positive. To illustrate this relation between higher goals and higher achievement, one can consider two students with similar characteristics (pretest score, math performance, etcetera) of which one student received the advanced performance goal (the highest goal, coded as a 5) and one student received the basic performance goal (coded as a 3). The difference between these two students on the posttest is (5 x 1.99 – 3 x 1.99 = 9.95 – 5.97 =) 3.98 which means that the student with the advanced performance goal is found to perform almost 4 points higher on the assessment than the student with the basic performance goal. Inclusion of this variable increased the fit of the model: the deviance decreased by 7.14, which is a significant improvement (p = .008; the critical value in a chi-square distribution with df = 1 is 3.84 for p = .05, as the models differ in 1 parameter).

When comparing the main effect model to the interaction model, the interaction term is found to be significant: both the interaction between goal and pretest as well as the main effect of the performance goal on the posttest are significant predictors in this model. Inclusion of this interaction effect led to a decrease in deviance of 4.86, which is a significant improvement (p = .027; again compared to the critical value of 3.84 as the models differ in 1 parameter). The interaction between goal and pretest for students' posttest results is illustrated in Figure 2. In this figure, the relation between the performance goal and the posttest is depicted for three types of students having differing pretest performance, namely 1) performance being one *SD* below the pretest average, 2) performance at the pretest average, and 3) performance being one *SD* above the pretest average. The upper part of the line for the student that initially performed the lowest (being 1 *SD* below the pretest average) and the lower part of the line for the student that initially performed the highest (1 *SD* above the pretest average) have been depicted as a dotted line rather than a continuous line in Figure 2 to indicate their implausibility: the incidence of an initially low performing student receiving an advanced goal and an initially high achieving student receiving a below minimum goal is considered improbable.

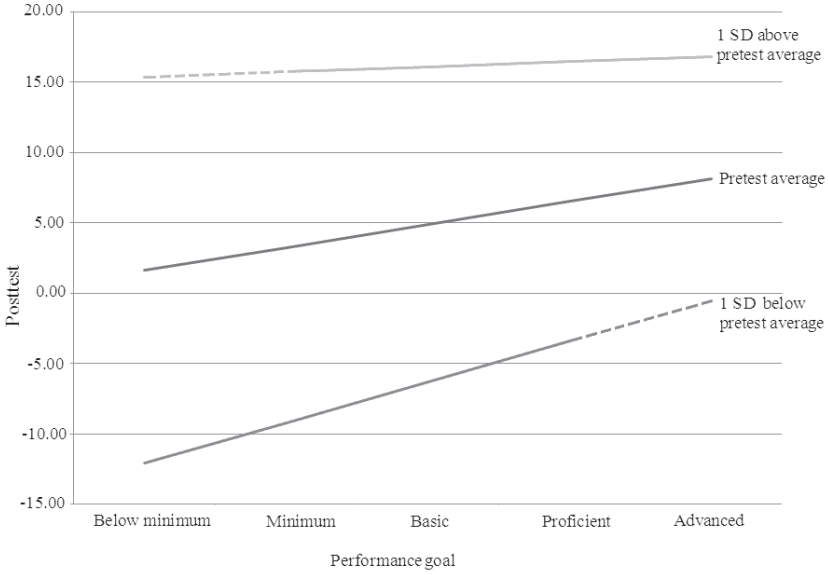


Figure 2. Interaction between the Performance Goal and the Pretest on the Posttest

From the results presented in Figure 2, it can be seen that there is a marked effect of higher performance goals for initially low performing students (being one *SD* below the pretest average). The relation between high goals and high posttest performance is also distinct for students that initially performed at the pretest average. For students with a pretest score of one *SD* above this average, this relation is still positive but less pronounced than for the other two types of students.

4.5 Conclusion and discussion

In this chapter, we investigated the relation between teacher-set performance goals and students' reading comprehension as the goals were a central part of a teacher Professional Development (PD) program targeting this subject area. The participating teachers set performance goals by selecting one of five performance categories for each of the students in their class. These categories (below minimum, minimum, basic, proficient, and advanced) pertained to test-score intervals on the end-of-the-year assessment which the teachers had defined themselves in a previous stage of the program.

For the first research question, we focused on the use of the multistep procedure which was developed to assist teachers in reflecting on and reconsidering their goals' appropriateness before

deciding on a final version. The difference between the initial goal (set at the beginning of the multistep procedure) and the final goal was tested and teachers' goal adaptation average was significantly larger than zero; demonstrating that there was significant change across this procedure. This change is viewed as an indication that the goals were set in a well-considered manner and viewed as support for using a multistep approach to goal setting.

For our second research question, we focused on teachers' attainment of the final student-specific performance goals as this was a focal point during the program. For 21 percent of students, their test results were lower than the test score interval which was selected for them. For 79 percent of students, their goals were attained by performing at the desired level or higher. It was found that 38 percent of students performed higher than their selected test score interval, which is considered a rather high percentage of students who surpassed their teacher-set goal. Perhaps these goals were set too low given these students' capabilities. Or perhaps the teacher set an appropriately challenging goal that benefitted performance to such an extent that the student attained a test score which was (slightly) higher than the selected test score interval. The way in which one should interpret these goal attainment results is not very straightforward, but it is facilitated by the results of the third research question. Here, we investigated the relation between the performance goals and students' reading comprehension. It was found that the teacher-set performance goals were a significant, positive predictor of students' performance on the posttest. In the analyses, we accounted for relevant covariates and demonstrated that higher goals were associated to higher performance - a finding which concurred with literature on goal setting (Locke & Latham, 1990; 2002) and teacher expectancies (Jussim & Harber, 2005; Rosenthal & Jacobson, 1968; Rosenthal, 1987).

For our fourth research question, we were interested in whether the effect of the performance goal on achievement depended on students' previous reading comprehension results; this was found to be the case. It was demonstrated that initially low-achieving students benefitted extra from high teacher goals; again, a result in line with previous findings (e.g., Good & Brophy, 2003). There was a limited impact of the performance goals for initially higher achieving students. We expect this to be due to a ceiling effect. If we would have used a larger number of performance categories, this would have increased the possibility to set a challenging goal for initially high achieving students. However, working with too many categories may have pragmatic constraints.

In this section, we would like to acknowledge several limitations to our study. First of all, we evaluated the deliberateness with which the teachers set their performance goals by assessing the degree of change across the multistep procedure. This evaluation did not include all the participating teachers: some teachers had not completed the required homework assignment. These latter teachers had participated in the goal setting meeting (including the step of team discussion) and had set final performance goals, but they had missed the previous steps of this

procedure. It remains unclear how we can establish the deliberateness of their final performance goals. Moreover, as previously acknowledged in the chapter, information on the degree of change across a certain procedure is a restricted way of evaluating the goals' deliberateness. Additional in-depth interviews or surveys on how the multistep procedure influenced teachers' goals would have complemented the findings of our study. A complicating matter is that the multistep procedure aimed to help set appropriate goals for students (being difficult but not too difficult given these students' capabilities), but we do not know what the most appropriate goals are.

Another limitation is related to the fact that the teacher-set performance goals were based on test score intervals. We assumed that the use of the standard setting procedure, in which teachers considered reading comprehension items to establish these categories, made these categories more tangible. Assessing how teachers experienced working with such goals and how they explain (a lack of) attainment, also in relation to the content of the entire PD program, would have yielded valuable information.

The third and last consideration pertains to the fact that, in the current study, the goals were not communicated to the students. Future studies in which these goals (or a simplified version of these goals) are communicated to the students, and in which students participate in the process of setting goals, are considered a worthwhile endeavor, as teachers and students then share the responsibility for the attainment of the goals.

Here, we would like to highlight an important finding of the current study. In the prediction of students' results, we investigated the effect of the performance goal variable while controlling for common predictors of students' reading achievement. For instance, we included prior achievement in reading, a proxy for general academic ability, and students' grade. Covariates such as educational level of the parents, Special Educational Needs, and students' sex were originally included but excluded from our models as they were found to be non-significant predictors of achievement. The results of this study empirically support the relation between the performance goal and student achievement which was presumed by the PD program. However, other student characteristics, not measured in this study - such as motivation, concentration, or effort -, might have played a role in teachers' goal setting decisions. More research is needed in order to better understand the relationship between performance goals and student achievement.

Working with goals is considered a promising approach as the goals direct the focus toward the attainment of desired outcomes, particularly when these goals are set at an ambitious level. More experimental research is needed to further our understanding on how teacher-set performance goals are causally related to student achievement. In the current study, we illustrated how goal setting can be implemented in a context which is aimed at performance improvement.

5. Exploring teacher implementation of the professional development program

Abstract: A teacher Professional Development (PD) program was developed aiming to improve students' reading comprehension by making instruction in this subject more goal-oriented, focused, clear, and better suited to students' needs. As part of this PD program, we trained teachers in the use of a) Direct Instruction, which is a teacher-centered model for instruction focused on the content and structure of a lesson, b) modeling, which is an instructional technique in which the teacher demonstrates how to solve a problem or how to apply a reading strategy by thinking aloud, and c) differentiation, which is an instructional practice in which the teacher attends to differences between students via the provision of extended instruction and by making adaptations in assignments students are expected to complete. In the current study, we focused on teachers' implementation of these three instructional practices. Comparing observation results collected at the start of the program to those collected at the end of the program, a significant improvement was found for the number of teachers who implemented modeling. Furthermore, students whose teachers modeled showed significant higher reading comprehension results in comparison to students whose teachers did not model, with an effect size of $d = .24$, 90% CI [$d = .03$; $d = .46$], though the number of teachers who implemented modeling on the postmeasurement was relatively small. The implementation of both Direct Instruction and differentiation was found to be rather limited and did not change from pre- to postmeasurement.

5.1 Introduction

Teacher professional development is a key mechanism to improve classroom instruction and, subsequently, student achievement (Borko et al., 2010; Cohen & Ball, 1990; Yoon et al., 2007). Many recent educational reforms rely heavily on teachers' implementation for their success (Borko et al., 2010; Desimone, 2009; Garet et al., 2001; Hill, 2007; OECD, 2005). As teachers need support and guidance to implement new teaching routines, Professional Development (PD) programs are commonly used to help teachers realize this desired change (Black & Wiliam, 1998b; Borko, 2004; Guskey, 2002). For the evaluation of the PD programs, information on the degree to which teachers have implemented the program - also referred to as the degree of *teacher change* - is essential. When one does not find any effect of a PD program on student achievement, this implementation data can give indications on why the program has not worked as it shows whether a) the teachers have not changed as expected, or b) the students have not changed after the teachers have changed in the desired way. When one does find effects on student achievement, this implementation data will indicate whether c) the program has worked as intended as both teachers and students have changed in the desired way, or d) other factors may have caused the observed effects on students (Wayne et al., 2008).

We developed a teacher PD program in which we trained and coached second- and third-grade teachers in the Netherlands (student age: approximately 7 to 9 years old). In Chapter 2, positive effects of teachers' participation in this program on student achievements were demonstrated. In the current chapter, we investigate to what extent teachers have changed their practices in accordance to the program. More specifically, we investigate the participating teachers' application of a) Direct Instruction, b) differentiation, and c) modeling, as these practices were explicitly targeted during the program. The implementation of these practices will be linked to students' reading results: if teachers with larger levels of implementation attain higher reading results than teachers with lower levels of implementation, this is empirical support for the mechanisms through which the program is expected to foster student reading comprehension.

The program was designed following the tradition of *applied research*; a research paradigm which aims to produce knowledge for the solution of a practical problem (McMillan & Schumacher, 1989). Its development followed recent concerns on students' performance in this subject area (e.g., Ministry of Education, 2008; 2010). First, these performance concerns are addressed, after which we continue to common reading instruction practices in the Netherlands. Subsequently, we discuss the theoretical background of the PD program in relation to the hypothesized change in instruction, before continuing to the details of the current study.

5.2 Theoretical framework

5.2.1. Concerns and current instructional practices for reading comprehension

The aforementioned concerns in the Netherlands were the result of Dutch students' performing lower than expected on both international and national assessments. For example, on the Dutch periodical assessment of educational achievement (known as PPON), 30 percent of the third-grade students read at a level which, according to reading experts and teachers, should be attainable for 75 percent (Van Berkel et al., 2007). Furthermore, although the scores on the 2011 international PIRLS assessment (targeting fourth-grade reading) indicate that, comparatively speaking, students in the Netherlands perform rather well, the average achievement of the Dutch students is significantly lower than in 2001 (Meelissen et al., 2012). The national performance concerns pertain particularly to the degree to which struggling, poorly performing readers are prepared for later schooling and the work force (Inspectorate of Education, 2007; 2010b). On the 2012 international PISA assessment (targeting - among other areas - the reading skills of 15-year olds), it is found that almost 14 percent of the Dutch students demonstrate such low levels of literacy that they are considered to have difficulties participating in society (Kordes et al., 2013). As "[r]eading is essential to our success in society" (Snow et al., 1998, p. 17) and the long-term effects of reading well at an early age have been widely established in the literature (e.g. Bodovski & Youn, 2011; Snow et al., 1998), the improvement of students' reading achievements is a priority for Dutch policymakers and practitioners, and thus its place on the research agenda is obvious.

Different causes have been suggested for the unsatisfactory reading results of Dutch students. For instance, several educational authorities attributed these insufficient results to the fact that, for schools and teachers, it was unclear what students should know and do at certain time points (Council of Education, 2007; Inspectorate of Education, 2011; Ministry of Education, 2010). Clearly defined performance goals were desired as these goals were assumed to make instruction more targeted which, subsequently, was assumed to result in improved student outcomes (Expert group Continuous Learning Progression, 2008). This line of reasoning was supported by findings from e.g., goal setting theory (Locke & Latham, 1990; 2002).

Another possible cause for the unsatisfactory results pertained to the quality of instruction which is provided in this subject area. In Dutch classrooms, reading comprehension lessons often take the following sequence: first, students read a text either out loud or in silence. Second, a few textbook questions about the text are discussed with the whole class, after which students have to answer the remaining questions independently (alone or in pairs). Last, the correct answers are discussed with the whole class (Aarnoutse, 1992). Teachers in the Netherlands are found to focus too much on asking students questions about a particular text at hand and they provide little explicit instruction on how students can improve their comprehension (Aarnoutse & Weterings,

1995; de Jager et al., 2002; Van Elsäcker, 2002). Similar results have been reported in other countries such as the United States, Norway, and Belgium (Andreassen & Braten, 2011; Liang & Dole, 2006; Van Keer & Verhaeghe, 2005). According to Collins-Block and Pressley (2002 in Houtveen, 2002), teachers do not offer much instruction in reading comprehension because they are unaware that this may improve comprehension. Via immersion, students are expected to become proficient readers on their own. Yet teachers should provide explicit instruction, in particular in the use of reading strategies (i.e., tools that can help the reader to better understand the text at hand), in order for students to attain relevant knowledge and skills which will benefit their comprehension of texts (National Reading Panel, 2000; Pressley, 1998; Snow, 2002).

A third possible cause for the unsatisfactory results of students pertains to the finding that Dutch teachers frequently struggle in meeting different students' needs. *Differentiation* is defined as "an approach to teaching in which teachers proactively modify curricula, teaching methods, resources, learning activities and student products to address the diverse needs of individual students and small groups of students to maximize the learning opportunity for each student in a classroom" (Tomlinson et al., 2003, p. 121). During the reading comprehension lessons in the Netherlands, teachers are found to rarely differentiate between students (Van Berkel et al., 2007; Van Elsäcker, 2002). When differences between students are targeted, this is done by differentiating in the assignments students are expected to make; 50 percent of teachers state that they differentiate in this aspect. Though differentiation in exercises is important, it is not sufficient as struggling readers need more explanation and instruction to keep up with their classmates (Houtveen & Van de Grift, 2012; Vernooy, 2005). Yet only 7 percent of teachers indicate that they differentiate through modifications in instruction; 90 percent of teachers state that they solely provide whole-group instruction when teaching reading comprehension (Van Berkel et al., 2007). Moreover, when differences between students are targeted via differentiation in instruction and assignments, most attention is paid to struggling readers. Very little attention is paid to students who read above grade level (discussed in Meelissen et al., 2012); again, similar to findings in, for example, the United States (Reis et al., 2011).

Last, it has been hypothesized – although not empirically researched – that Dutch primary school teachers find reading comprehension a difficult subject to teach due to the complexity of the reading comprehension skills and the inadequacy of the curricular textbooks used in Dutch primary schools (Droop et al., 2012; Houtveen & Van de Grift, 2012; Stoeldraijer & Forrer, 2012). These textbooks have been criticized as being "more bulky than necessary, containing a substantial amount of material that has little or nothing to do with learning to read" (Houtveen & Van de Grift, 2012, p. 88). They also contain a large number of reading strategies, but not all of these strategies which are presented as "effective" can be supported by empirical evidence (Droop et al., 2012; Stoeldraijer & Forrer, 2012). The inadequacy of the curriculum is considered

to be problematic as teachers in the Netherlands are known to follow the curricular textbooks to a very large extent (Meelissen et al., 2012).

Overall, support had been gained for the view that Dutch teachers' reading comprehension instruction can be improved on various aspects. For this purpose, a multicomponent teacher PD program was developed.

5.2.2. A multicomponent teacher PD program

We set out to develop a program that would improve students' reading results by making teachers' instruction more goal-oriented, focused, clear, and better suited to students' needs. For this purpose, we developed the PD program which contained three components, namely 1) setting standards and performance goals for every student, 2) applying formative assessment and data use, and 3) acquiring relevant instructional skills and (content and curriculum) knowledge in reading comprehension. As part of these components, we discussed the instructional practices of Direct Instruction and modeling during after-school meetings and we provided suggestions how to improve teachers' differentiation practices. We will refer to the three components by referring to the questions of 1) *Where am I going?*, 2) *How am I going?*, and 3) *How can I improve how I am going?* The relation between the components and these questions has been discussed in Chapter 1.

We aimed to stimulate goal-oriented instruction which would be better suited to students' needs by asking teachers to set student-specific performance goals and by monitoring students' progress in relation to these goals (components 1: *Where am I going?* and 2: *How am I going?*). With help of these two components, it was expected that teachers would attend more to differences between students and that differentiation would be fostered as a result. Furthermore, we aimed to stimulate focused and clear instruction by the implementation of Direct Instruction and modeling, as well as by training teachers to become more knowledgeable in the important determinants of reading comprehension skills, key concepts and the curriculum (component 3: *How can I improve how I am going?*). Moreover, when teachers are more knowledgeable in reading comprehension, this is expected to benefit their differentiating practices as the teachers can then more easily detect which students do not master the essential skills and knowledge – for these teachers, it becomes (more) evident which students are in need of additional support. Hence, the components must be seen as fostering behaviors in an inter-related manner. This is why we considered the *molar* approach of Chapter 2 (in which we investigated the overall relationship between the entire program and student achievement) to be an appropriate approach for the evaluation of the effects of the program.

Direct Instruction is a teacher-centered model for instruction focused on the content and structure of a lesson; it is an effective instructional practice (Borman et al., 2003; Muijs & Reynolds, 2011), particularly for struggling readers (Houtveen & Van de Grift, 2007; 2012).

There are slight variations in the literature on which activities the Direct Instruction model entails (different models are discussed in e.g., Borman et al., 2003; de Jager, 2002; Houtveen & Van de Grift, 2007; Muijs & Reynolds, 2011; Veenman, Leenders, Meyer, & Sanders, 1993). This variation is also acknowledged in research of Baumann (1988) who describes that the definition of Direct Instruction "sometimes denotes the use of regimented, scripted lessons; other writers use the term to refer to a generalized set of teacher behaviors and classroom conditions related to high levels of student achievement" (p. 712). Commonly, the Direct Instruction model contains the following steps:

a) review and activation of the preceding subject matter,
b) presentation and explanation of the new subject matter,
c) guided practice and coaching,
d) seat work,
e) recapitulation of the current subject matter, and
f) preview of the subject matter to be addressed in the following lesson (Leenders, Naafs, Oord, & Veenman, 2010).

In the presentation of this model during our PD program, we stressed the elements pertaining to the *beginning* of a lesson in which teachers should review and activate previously discussed content, and give a clear account of the current lesson's objectives (abovementioned as steps a and b). We also stressed elements of the Direct Instruction model pertaining to the *end* of a lesson, where teachers should recapitulate the current lesson's objectives (step e), and provide a preview of the subject matter to be addressed in the following lesson (step f). We focused particularly on these elements because they capture the core of the lesson at hand, and focus both the teacher's and students' attention to the most important skills and knowledge the lesson aims to address. In a study conducted by the Inspectorate of Education (2010), only 40 percent of primary school teachers explicated the lesson's objectives at the beginning and at the end of a lesson, while other elements of the Direct Instruction model were implemented by a vast majority of teachers. Implementation of these Direct Instruction elements is assumed to make instruction more focused.

After knowing which objectives a teacher should focus on with his students, it is important to understand how to instruct students in such a way that they grasp the content at hand (useful for steps b, c, and d in the aforementioned Direct Instruction model). For this purpose, we discussed the practice of modeling. Modeling is demonstrating how to solve a problem by thinking aloud and linking the solution to skills or knowledge that the students already possess. It is an effective instructional technique (Fisher et al., 2008; National Reading Panel, 2000) and thought to be the best way through which teachers can demonstrate to their students how a reader interacts with a text (Taylor & Pearson, 2002 in Fischer, Frey & Lapp, 2009). When thinking aloud, teachers can

show students which knowledge, skills or strategies are appropriate, why it is important to use them and how to use them. Modeling in combination with Direct Instruction has been identified as an effective procedure to help struggling learners and for remediating learning disabilities, as found in the meta-analyses of Swanson and Hoskyn (1998). Authors such as Baumann (1988) and Houtveen and Mijs (2004) discuss modeling as part of the Direct Instruction model, as Direct Instruction involves "teachers showing, telling, modeling, demonstrating, explaining, teaching how various reading skills, processes, and strategies function" (Baumann, 1988, p. 714). For Muijs and Reynolds (2011), modeling is not part of the Direct Instruction model as such, yet it is viewed as a relevant instructional approach that teachers should apply when they want to provide an effective Direct Instruction lesson. Teachers in Dutch primary schools are still rather unfamiliar with modeling, although this instructional approach has received some attention in journals targeting teachers and schools rather recently (e.g., Filipiak, 2006; Loman & Marreveld, 2010) and in other reading improvement PD programs (for example, Droop et al., 2012). Through the implementation of modeling, we expect instruction to become clearer.

When the majority of students are working independently on their assignments during the seat work phase of their lesson (step d in the Direct Instruction model), the teacher can provide additional small-group instruction and thus differentiate between students. This extended instruction for students in need is strongly recommended to help struggling students in their attainment of relevant skills (Good & Brophy, 2003). Teachers can also differentiate in the assignments students are expected to complete (Tomlinson et al., 2003). Several authors associate differentiation to the Direct Instruction model as well (e.g., Becker, 1977 in Baumann, 1988; Houtveen & Mijs, 2004). For our PD program, we particularly see the implementation of differentiation as a way to attain the performance goals. Teachers should attend more to different students' needs and modify their teaching in such a way that different goals for different students will be attained. By implementing differentiation, we expect teaching to become more goal-oriented (as it focuses on the attainment of the performance goals) and better suited to students' needs.

5.2.2.1 Stimulating implementation during the PD program

To help teachers implement Direct Instruction, modeling, and differentiation, they received training and constructive feedback on their implementation on several occasions. The PD program was conducted in the school year of 2011-2012, and relatively at the start of the program we conducted lesson observations (Sept. - Dec. 2011) to attain information on the degree to which the teachers already applied the instructional practices under study[22]. Immediately after these observed lessons, we discussed the results with the teachers and provided them with constructive feedback in relation to these three practices.

[22] From September to December, these instructional practices were not discussed during the PD program's meetings.

The instructional practices of Direct Instruction, modeling, and differentiation were also discussed with the participants during several of the after-school meetings which were held throughout the course of the program. In the case of modeling, specific feedback on teachers' own implementation was provided during the meeting in which we discussed this practice (as teachers were asked to model "on the spot", and constructive feedback with respect to their application was provided instantly). In the case of Direct Instruction, this practice was discussed during two meetings, and just before the second meeting on this topic, teachers received feedback on their own implementation (provided by the school's principal or internal support coordinator after the first meeting on Direct Instruction). Both modeling and Direct Instruction were discussed in meetings targeting the third component (*How can I improve how I am going?*). For more details on these meetings, conducted in February and in April/May of 2012, one is referred to Appendix 1 of this dissertation. For differentiation, a slightly different approach was used. As part of the first component of our program (*Where am I going?*), teachers were asked to set performance goals for each individual student. We frequently discussed these performance goals during the meetings in which recent student performance was connected to these goals: this was part of the program's second component (*How am I going?).* As part of these meetings on the second component, we gave prompts and hints how to actively target these goals, e.g., by providing extended instruction to students with certain performance goals or by differentiating in assignments. These suggestions were also provided during meetings on the third component (*How can I improve how I am going?*). Considerable attention was paid to adequately dealing with differences between students throughout the course of our PD program.

The participating teachers were observed again at the end of the school year (May to July of 2012), prior to the program's last after-school meeting. Similar to the beginning of the school year, they were provided with constructive feedback on their implementation immediately after this observation. The observation results were also discussed during the program's final meeting, in which the program was evaluated by the participants and the researchers.

5.2.2.2 Studying implementation

We expected that students' reading comprehension performance would improve when instruction in this subject would become more goal-oriented, focused, clear, and better suited to a student's needs. For this purpose, the instructional practices of Direct Instruction, differentiation, and modeling were discussed. The implementation of these practices is assumed to be stimulated through the training we provided in these practices and fostered by other relevant knowledge and skills we addressed during the program. Hence, the implementation of these practices is assumed to be the result of the multicomponent design of the program.

In Figure 1, our theory of action is provided. In this figure, we made use of Wayne, Yoon, Zhu, Cronen and Garet's (2008) distinction between the *theory of teacher change* in which the

content of the PD program is linked to change in teachers' practice, and the *theory of instruction* in which the changed practice is linked to change in students' performance. According to Wayne et al. (2008), both theories are necessary to understand how professional development works.

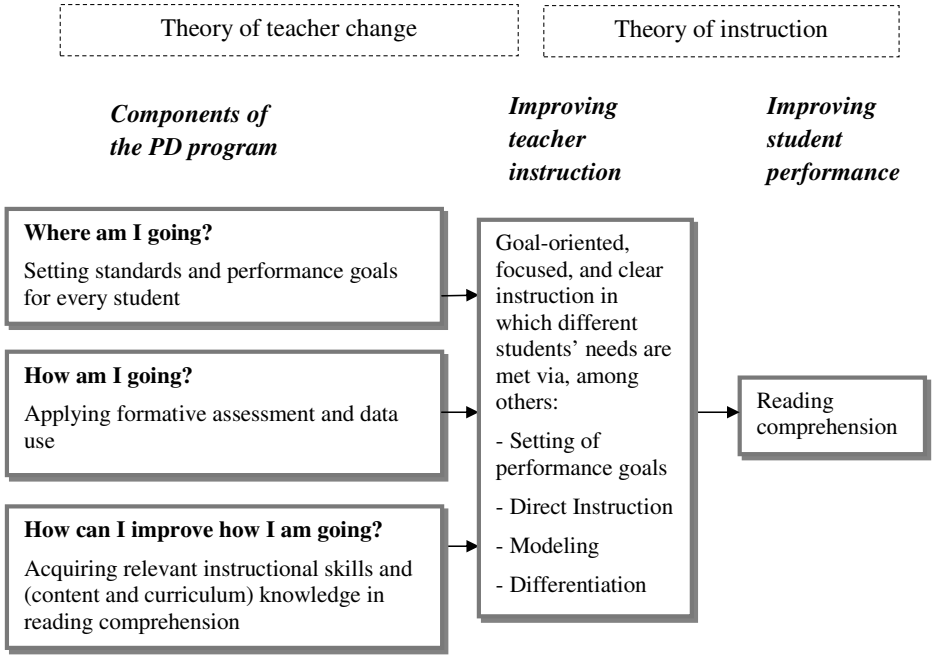


Figure 1. Theory of action

5.2.3. The current study

In this study, we investigate teachers' hypothesized improved implementation of a) Direct Instruction, b) modeling, and c) differentiation which were targeted in a multicomponent teacher PD program. The data on teachers' implementation will be linked to student achievement to assess the effect of stronger implementation. In addition, as some of the instructional practices are known to be particularly beneficial to struggling readers, we are interested whether teachers who implemented these practices to a larger extent attained better results with these specific students in comparison to teachers who implemented these practices to a lesser extent.

In this chapter, the following research questions are addressed:

1) Have teachers improved in their application of Direct Instruction, modeling, and differentiation after following a teacher professional development program targeting these practices?

2) Are differences in the implementation of these instructional practices related to students' reading comprehension?

3) Does the effect of increased implementation of these practices on achievement depend on students' initial performance?

5.3 Method

In this study, pretest posttest designs were used to investigate the research questions at hand.

5.3.1. Participants

A total number of 19 schools and 33 teachers from the northern part of the Netherlands participated in our PD program. Schools participated in this study on a voluntary basis: no financial or other compensation was provided. In the current study, we only included those teachers who were included in the effect study of the program - see Chapter 2 - and for whom the observation data was complete. This resulted in a sample of 24 teachers[23]. The average number of years of experience of the 24 teachers that participated was 14.09 years (*SD* = 11.75). The vast majority of our participating teachers were women; three participants (12.5 percent) were men.

5.3.2. Instruments and variables

5.3.2.1 Instruments and variables used to assess teachers' implementation of the instructional practices under study

Observation instrument: The observation instrument, used in this study to address the first research question, was an amended version of the observation instrument used by Kooiman et al.

[23] In total, 29 teachers were included in the analyses as discussed in Chapter 2. Here, we elaborate on the five cases that were not included in the analysis on changes in instructional behavior (this chapter). For one teacher, we could not collect any observation data due to difficulties in scheduling an observation appointment. Two teachers who taught reading comprehension dropped out of the program due to burn-out and retirement during the spring of 2012, hence their observation results were not complete. One teacher of a multi-grade class (containing third and fourth grade) was observed teaching reading comprehension to the fourth-grade class while she was asked to teach the third grade class. As a result, her instruction toward the third-grade students was very limited as these students worked on independent seatwork the entire lesson. The data which were collected with help of the observation instrument did not provide a realistic portrait of the teacher's instructional behavior and were therefore excluded. One teacher could not be observed teaching whole-class reading comprehension as this part of the reading comprehension curriculum was provided by another teacher (who also participated in the PD program).

(2005) in which a low-inference measure (i.e., time-sampling) and a high-inference measure were combined. This instrument (see Appendix 7) was used to assess teachers' degree of implementation (change from pre- to postmeasurement) for the purpose of evaluating the program as well as to provide the research team members with input in order to give the teachers constructive feedback with respect to their implementation of the practices under study.

The high-inference measure contained 16 items pertaining to different aspects of teachers' behavior which were filled in directly after the lesson was observed. All items had dichotomous response options, with a 0 (*no*) or a 1 (*yes*) depending on whether or not the implementation of these practices was observed. In this chapter, we focus on the items on the high-inference measure which pertained to the implementation of Direct Instruction, modeling, and differentiation. The low-inference measure focused on different aspects of the teachers' and students' activities, which were coded every 2 minutes. In the current study, we discuss the results of the low-inference measure pertaining to teachers' differentiating practice only. The high-inference measure and the low-inference measure each targeted different aspects of differentiation: the former focused on the differentiation in assignments that students needed to complete, while the latter focused on extended instruction. The inter-rater reliability (Cohen's kappa) between the researchers and research assistants who conducted the observations was considered satisfactory for both the high-inference and low-inference measure (for the measurement at the beginning of the school year, $k = .83$ and $k = .74$, and for the measurement at the end of the school year, $k = .85$ and $k = .82$). Using this instrument, observation data was collected in September to December of 2011[24]; these observations are referred to as the *premeasurement* for the remainder of this chapter. In May to July of 2012, observations were conducted as well; these observations are referred to as the *postmeasurement*. Here, we will discuss the separate variables as used in our analyses.

Direct Instruction: The application of the Direct Instruction model was measured using four items on our high-inference measure. During the PD program, we explicitly targeted the beginning and end of a lesson of the Direct Instruction model, and these elements were measured with help of the following items: 1) the teacher summarized the content of the prior lesson or activated relevant prior knowledge, 2) the teacher explicated the learning goal, content and/or topic of that lesson, 3) at the end of the lesson, the teacher returned to the learning goal of that lesson and/or the new skill/knowledge that had been addressed, 4) the teacher connected the content of the current lesson to the following lesson. All items contained the dichotomous response options of 0 (*no*) and 1 (*yes*). In the analyses, the Direct Instruction-variable pertained to

[24] For two teachers from the same school, the premeasurement observations were conducted in the beginning of May 2012, prior to the seventh PD program meeting, due to scheduling difficulties. As these observations were conducted prior to the most vital meetings on instructional behavior (being meeting 7 and 8), we included this data in the analyses. Postmeasurement observations at this school were conducted in July.

the sum score on these four items: its results were placed on an ordinal scale ranging from 0 (*none of the elements implemented*) to 4 (*all elements implemented*).

Modeling: To assess the application of modeling, we used one item on our high-inference measure, being whether the teacher modeled his/her application of knowledge, skill or strategy by thinking aloud. Teachers received a 0 (*no*) or a 1 (*yes*) with respect to their implementation of modeling.

Teachers' differentiating behavior – differentiation in assignments: One aspect of how teachers could meet different students' needs was by differentiating in the assignments that students were expected to complete. This was measured on our high-inference measure using two items: 1) the teacher differentiated for weaker students in the assignments that these students were expected to complete, and 2) the teacher differentiated for well achieving students in the assignments that these students were expected to complete. The variable we used in our analyses was constructed by taking the sum score on these items: as the two items were answered with either a 0 (*no*) or a 1 (*yes*), this sum score could range from 0 (*teachers did not differentiate in their assignments for students*) to 2 (*teachers differentiated in their assignments for both weak and well achieving students*).

Teachers' differentiating behavior – extended instruction (percentage of the lesson): A second aspect of how teachers could differentiate instruction to meet different student needs was by providing extended instruction to certain students. The current variable represented the percentage of the lesson spent on extended instruction. Using our low-inference measure, we coded the phase of the lesson every two minutes in which we distinguished between a) whole-class instruction, b) extended instruction, in which the teacher provided an individual student or a small group of students with additional instruction and/or supported them while they were completing their assignments (via questioning and scaffolding), while the rest of the class was working independently, or c) seatwork, where all students worked alone, in pairs, or in small groups. During the seatwork-phase, teachers frequently walk around in the classroom to ensure that students remain on task and to check for students' understanding; this is known as "making the rounds" (Good & Brophy, 2003, p. 314). The teachers might have provided assistance to students who had questions or who were struggling to complete the assignments during this phase, but the difference between extended instruction and help-during-seatwork pertains to, among other things, whether this additional help and instruction for students was scheduled in advance. Extended instruction is teacher-initiated instruction aimed at assisting students who are known to struggle and commonly follows whole-class instruction immediately. It can be signaled by the teacher through the use of statements such as "Jim and Mary, please come to my desk while the rest of the class can start making their assignments now".

5.3.2.2 Instruments and variables used to predict student attainment

The instruments and variables used for the second and third research question have also been used in Chapter 2 and Chapter 4. Here, we will resume them briefly.

Reading comprehension assessment results: To measure the students' reading comprehension skills, we used the Cito standardized reading comprehension assessments which were developed by the Netherlands Institute for Educational Measurement. In these reading comprehension tests, students are asked to read several texts and answer multiple-choice questions referring to the word, the sentence, and the text levels. The reading comprehension assessment results were registered on a continuous scale, and this scale ranged from -87 to +147 (end of grade 1 to mid-term grade 6). The negative symbol (-) for a large part of the assessment scale should not be interpreted as having a negative connotation; -87 is simply the (arbitrary) starting point of this scale. The assessment results in June 2012 were used as the posttest data. The assessment results of the June 2011 assessment were used as pretest data, which were controlled for in the analyses (this variable had been grand-mean centered to facilitate its interpretation in the analyses).

Mathematics assessment results: We wanted to include a proxy for general academic ability and added information on the students' prior mathematics performances in the analyses. The mathematics results were collected on the Cito standardized mathematics assessments. The assessment results were registered on a continuous scale which ranged from 0 to 169. The June 2011 grade-specific mathematics assessment results were used as a covariate in the analyses (grand-mean centered in the analyses to facilitate its interpretation).

Sex: Boys were the reference group for this dummy-coded variable.

Grade: For this dummy-coded variable, second grade was the reference group.

Teacher-set performance goal: Teachers that participated in our program set a performance goal for each of their students by selecting one out of five performance categories (below minimum, minimum, basic, proficient, and advanced). This variable thus ranged from 1 (*below minimum*) to 5 (*advanced*). As we identified that the performance goal was a significant predictor of achievement in Chapter 4, it was included as a covariate in the current analyses.

Multi-grade classroom: This dummy-coded variable, with single-grade-classrooms as the reference group in the analyses, was a classroom characteristic which was controlled for in the analyses as we identified that this variable was a significant predictor of achievement in Chapter 4.

5.3.3. Analyses

For the research question on teachers' implementation of Direct Instruction, differentiation, and modeling, we investigated changes between the observations results collected in the fall of

2011 and in the summer of 2012. Changes on the following four variables were analyzed: 1) the sum score of the items on the high-inference measure pertaining to Direct Instruction, 2) the score on the item on the high-inference measure pertaining to modeling, 3) the sum score of the items on the high-inference measure pertaining to differentiation in exercises, 4) the percentage of time spent on extended instruction. As we expected teachers' instruction to improve after participating in our PD program – i.e., an increase in (sum) scores and an increase in the percentage of time spent on extended instruction -, we conducted one-sided statistical tests for these paired data. Due to our sample size (n = 24) and accompanying limited degree of statistical power, we set the significance level at $\alpha = .10$. As we aimed to conduct multiple statistical tests in this study (i.e., a test for each of the four variables), we applied a Bonferroni correction to adjust the alpha level for chance capitalization. The alpha was set at $\alpha = .10 / 4 = .025$. For normally distributed data, the paired samples t-test was appropriate. Preliminary analyses of the data for several of the variables showed strongly skewed distributions. In this situation, we used the (one-sided) non-parametric Wilcoxon signed rank test which evaluates systematic differences within pairs: it tests whether or not the median of the difference scores equals zero (Siegel, 1988).

For the variables for which we found significant differences between the fall measurement and the summer measurement - indicating change after following the program -, we investigated the effects of implementation on students' reading results (our second research question). It was analyzed whether the performance on the posttest was significantly higher for students whose teachers demonstrated a large degree of implementation in comparison to students whose teachers demonstrated a lower degree of implementation while controlling for the aforementioned covariates; this would be considered as empirical support for the mechanisms through which the program was expected to foster student reading comprehension. Again, we conducted one-sided statistical tests as we expected higher levels of implementation to be beneficial to achievement. Given our sample size and accompanying limited degree of statistical power, we set the significance level at $\alpha = .10$ for this analysis as well. In order to identify whether the effect of implementation on achievement depended on students' initial achievement - the focus of our third research question -, we added an interaction term between implementation and pretest performance in the statistical model which had been used to analyze the second research question. As we expected weaker students to benefit from increased levels of implementation, we again conducted one-sided hypothesis testing (using $\alpha = .10$). A multilevel regression analysis was performed with the help of *MLwiN* software (Rasbash, Browne, Healy, Cameron, & Charlton, 2011), with students at level one and teachers at level two.

5.4 Results

5.4.1. Teachers' implementation of Direct Instruction

For the implementation of Direct Instruction (focusing on the beginning and end of a lesson), the results are presented in Table 1. In this table, we present the frequencies of each of the possible sum scores to illustrate the distribution of this variable.

Table 1

Summary of Teachers' Implementation of Direct Instruction at Pre- and Postmeasurement

		Direct Instruction[a]				
	Teachers	0	1	2	3	4
Observation	*Total n*	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)
Premeasurement	24	0	2 (8)	13 (54)	6 (25)	3 (13)
Postmeasurement	24	1 (4)	4 (17)	10 (42)	9 (37)	0

[a] The presented values are sum scores: higher scores indicate that more elements of Direct Instruction are implemented by the teacher.

From the results presented in Table 1, it can be seen that the results are quite comparable across the measurement occasions. There are some slight differences in the implementation of the Direct Instruction model as observed at pre- and postmeasurement. The scores attained in the summer of 2012 were slightly lower (with one teacher not implementing any Direct Instruction elements, and the highest attained score being a 3) in comparison to results attained in the fall of 2011. A median value of 2 was found at both occasions.

When taking a closer look at the different elements of the Direct Instruction model, it was found that teaching behaviors pertaining to the *beginning* of the lesson (i.e., the activation of relevant prior knowledge and explication of the lesson's learning objectives) were implemented most frequently. The Direct Instruction elements pertaining to the *end* of a lesson (i.e., whether the teacher returns to the lesson's objective at the end of that lesson and connects it to the content of the following lesson) were rarely observed. As we hypothesized to find an increase in sum scores after following the PD program, the findings presented above went against our hypothesis that the implementation of Direct Instruction would be higher in the summer of 2012 than in the fall of 2011, and there was no need to conduct a significance test.

5.4.2. Teachers' implementation of modeling

With respect to the implementation of modeling, two teachers were found to implement this instructional practice during the reading comprehension lessons observed in the fall of 2011. Nine teachers implemented modeling during their reading comprehension lesson in the summer of 2012; two of these nine teachers were the ones who also modeled at the premeasurement. The increase in the number of teachers who implemented modeling at the end of the school year in comparison to the beginning of the school year was substantial, and was statistically significant using the one-sided Wilcoxon signed rank test (p=.004).

5.4.3. Teachers' implementation of differentiation

For the observation results pertaining to teachers' differentiating behavior, we first discuss the results for differentiating in exercises which was assessed using our high-inference measure. When comparing the premeasurement to the postmeasurement for this sum score on differentiation in assignments (based on two items), we found no differences between these two occasions. None of the teachers differentiated in the assignments the weaker achieving students were expected to complete, at either measurement occasion. Only one teacher was found to differentiate in the assignments the well achieving students were expected to make, and this teacher did so at both pre- and postmeasurement. As there was no change in implementation from fall to summer, this went against our hypothesis of increased implementation and there was no need to conduct a significance test.

Next, we discuss the degree to which teachers[25] provided extended instruction, assessed using our low-inference measure. In Table 2, a descriptive overview is provided of the results on this measure.

[25] Two of our 24 teachers taught a multi-grade class containing both second and third grade students, and at the two measurement occasions, different grades were taught (meaning that, for example, at the beginning of the school year they were observed teaching second grade, and at the end of the school year, they were observed teaching third grade). We included these teachers' data in the analyses as extended instruction in reading comprehension is relatively uncommon in the Netherlands (Van Berkel, Krom, Heesters, Van der Schoot, & Hemker, 2007), and prior to the observations, it was considered likely that those who implemented this type of instruction would do so for both grades.

Table 2

Summary of Teachers Providing Extended Instruction at Pre- and Postmeasurement

Observation	Teachers *Total n*	Teachers providing extended instruction *n*	Percentage of lesson spent on extended instruction[a] *M* (*SD*)	Min.	Max.
Premeasurement	24	8	32.5 (16.9)	20.0	70.6
Postmeasurement	24	6	28.7 (12.3)	9.7	45.2

[a] These percentages pertain only to the lessons in which extended instruction was actually implemented.

It can be seen from the results presented in Table 2 that eight teachers provided extended instruction during the reading comprehension lessons which were observed in the fall of 2011, and six teachers did so during their reading comprehension lesson in the summer of 2012. Only three teachers provided this type of instruction at both occasions. For those teachers who did provide extended instruction, we observed quite some differences between teachers in the percentage of time they spent on this type of instruction. These differences were smaller at postmeasurement than at premeasurement. As these findings went against our hypothesis that teachers would increase their implementation of this type of instruction (i.e., increasing the degree to which they met different student needs and thus more teachers spending more time on extended instruction), there was no need to conduct a one-sided statistical test.

5.4.4. Predicting student achievement

Next we investigated whether students whose teachers implemented the instructional practice of modeling on the postmeasurement (the only indicator for which we found a significant improvement from pre- to postmeasurement) attained higher posttest results. In Table 3, descriptive data are presented for the variables which were used in the multilevel analyses. Here, we used the results of 332 students who were taught by 23 teachers[26]. Already in an early stage of the analyses, we found that students' grade was a non-significant predictor of performance. As it is not a significant predictor of achievement, we do not distinguish between the test results of second- and third grade students as presented in Table 3. Yet we included this variable in the models presented in Table 4 (further on) to be consistent with the models as presented in the previous chapter of this dissertation.

[26] One teacher did not set performance goals for his students, and therefore his class (containing 9 students) was excluded from this analysis as we included this significant predictor of achievement.

Table 3

Student and Classroom Characteristics

	M	*(SD)*	*n*	(%)
Student level characteristics				
Girls			177	(53)
Boys			155	(47)
Performance goal: below minimum			6	(1)
Performance goal: minimum			45	(12)
Performance goal: basic			91	(25)
Performance goal: proficient			128	(35)
Performance goal: advanced			62	(17)
Mathematics performance	55.04	(17.25)		
Pretest	12.14	(18.11)		
Posttest	28.33	(16.84)		
Classroom characteristics, at the student level				
Grade 2			181	(55)
Grade 3			151	(45)
Single-grade			168	(51)
Multi-grade			164	(49)

The results of the multilevel analyses, in which the effect of teachers' modeling behavior on reading comprehension achievement was estimated, are presented in Table 4. The first model, called the start model, contained the intercept and the covariates at the student and the classroom level but did not include the variable modeling. In the second model, called the main effect model, this variable was included (dummy-coded, with the group of teachers who did not model at the postmeasurement being the reference group). In this way, we could analyze whether or not this variable added value when predicting student achievement. In the third model, called the interaction model, we investigated the interaction between modeling and the reading comprehension pretest to assess possible differential effects. All models presented below contain unstandardized coefficients.

Table 4

Multilevel Models Predicting Achievement in Reading Comprehension

	Models					
	Start Model		Main effect model		Interaction Model	
Predictors	Coeff.	*S.E.*	Coeff.	*S.E.*	Coeff.	*S.E.*
Fixed Part						
Constant	17.05*	3.28	16.52*	3.22	16.85*	3.21
Girl	2.60*	1.19	2.69*	1.18	2.60*	1.18
Grade 3	-1.13	1.66	-1.25	1.58	-1.10	1.58
Math performance	0.15*	0.05	0.15*	0.05	0.15*	0.05
Pretest	0.54*	0.05	0.53*	0.05	0.58*	0.06
Multi-grade classroom	4.04*	1.37	3.38*	1.33	3.69*	1.34
Performance goal	2.37*	0.75	2.36*	0.74	2.25*	0.74
Modeling			2.47[a]	1.34	2.51[a]	1.33
Modeling x pretest					-0.11[a]	0.07
Random Part						
Variance at classroom level	2.67	2.90	1.30	2.44	1.28	2.42
Variance at student level	103.26	8.27	103.35	8.28	102.55	8.21
Deviance	2488.86		2485.80		2483.20	
No. of teachers	23		23		23	
No. of students	332		332		332	

* $p < .05$, two-sided.
[a] $p < .10$, one-sided.

When comparing the start model to the main effect model, one can see that modeling is in fact related to a significant higher performance on the posttest. Inclusion of this variable increased the fit of the model (the deviance decreased by 3.06, while the critical value in a chi-square distribution with $df = 1$ is 2.71 for $p = .10$, as the models differ in 1 parameter). When comparing

the main effects model to the interaction model, the interaction term is found to be a significant predictor of achievement but the fit of the model did not improve significantly (the deviance decreased with 2.6, while the critical value is 2.71 as again the models differ in 1 parameter). Hence, we will use the main effects model as our final model.

For the effect of modeling on student achievement, we observed an effect size of d = .24 (calculated by dividing its regression coefficient by the square root of the unexplained variance at the student level[27], using the coefficient in the main effect model), 90% CI [d = .03; d = .46]. According to Cohen's interpretation (1988), a value of d = .24 is a small effect.

5.5 Conclusion and discussion

In this study, we focused on teachers' application of three instructional practices, a) Direct Instruction, b) modeling, and c) differentiation, which were targeted in a multicomponent teacher Professional Development (PD) program aimed to foster students' reading comprehension achievement. Observation data, collected in the fall of 2011 and in the summer of 2012, were compared to investigate the hypothesized change in teacher instruction. This implementation data was linked to students' assessment results, in order to investigate whether teachers with larger levels of implementation attained higher reading results than teachers with lower levels of implementation. If teachers with larger levels of implementation attained higher reading results than teachers with lower levels of implementation, this was considered as empirical support for the mechanisms through which the program was expected to foster student reading comprehension.

In the case of modeling, we found a significant increase in teachers' implementation from pre- to postmeasurement. When using modeling as a predictor of student attainment, significant higher results were found for students whose teachers modeled in comparison to students whose teachers did not model. Yet it must be acknowledged that the number of teachers implementing modeling in the summer of 2012 was - albeit significantly larger than in the fall of 2011 – relatively small; nine out of 24 teachers demonstrated this behavior on the postmeasurement (of which two teachers already did so at premeasurement). In addition, the interaction effect between modeling and students' pretest performance, the focus of our last research question, did not significantly improve the fit of our statistical model.

The implementation of both Direct Instruction and differentiation appeared to be rather limited and did not change from pre- to postmeasurement. However, for both instructional practices, we must acknowledge the limitations in our observation instrument. In our design of the instrument, we focused on measuring only a very basic level of implementation of the

[27] This is an application of Cohen's (1988) formula of $d = \{\bar{x}(exp) - \bar{x}(control)\}/\sigma$ to a multilevel setting, for which we are interested in the variation within groups (i.e., level one).

behaviors we expected to see, as this was considered sufficient to help us provide constructive feedback to teachers on their implementation. We thus adapted an existing observation instrument in order to make the coding system as simple as possible, as we coded live behaviors (rather than video-data) and wanted to use a high-inference and a low-inference measure simultaneously. As a result, dichotomously scored items were used on our high-inference measure to simplify the measurement of implementation of Direct Instruction and differentiation (and, also, modeling). Differentiation was also assessed using the low-inference measure, in which we focused on whether or not teachers provided extended instruction. Summarizing, the data collected on this instrument and investigated in this chapter focused only on the *occurrence* of certain teacher behaviors. The participating teachers might have improved their implementation of Direct Instruction or differentiation in other ways than measured (for instance, for Direct instruction: by addressing the beginning of the lesson more elaborately in the summer of 2012 than in the fall of 2011. For differentiation: by differentiating in the types of questions students receive during whole-class instruction).

Despite these limitations in our observation instruction, we gathered valuable information on the implementation of targeted behaviors as it was found that very few teachers attained high scores on the indicators of Direct Instruction and differentiation. This is in line with other research findings in the Netherlands. For example, rather comparable findings on the implementation of Direct Instruction after following a teacher PD program – namely that the final part of the lesson being implemented poorly – were reported in the study of de Jager, Reezigt and Creemers (2002). With respect to differentiation, the limited degree of differentiation has been acknowledged in various publications (Inspectorate of Education, 2008; 2013; Van Berkel et al., 2007) and this is considered a rather complex skill for primary school teachers to attain (van de Grift, van der Wal, & Torenbeek, 2011). Directly after we conducted our lesson observations at postmeasurement, we asked teachers about this limited degree of implementation. For the Direct Instruction elements pertaining to the end of a lesson, teachers stated they experienced time constraints as they needed to continue to the next subject. For the differentiation elements, teachers referred to the limited guidelines in the curricular textbooks they used. In the Netherlands, it is known that teachers follow the content of the curricular textbooks to a very large extent in their lessons (Meelissen et al., 2012). Perhaps our program was not sufficiently intensive and not practical enough to facilitate change in these aspects. Research on the implementation of innovations has demonstrated that it is not easy to change teacher behavior (e.g., Fullan, 2001; Garet et al., 2008). Yet we suspect that also other factors might have opposed implementation. For instance, during the after-school meetings we experienced that teachers' response toward modeling was much more positive than toward Direct Instruction. While the former was considered new and interesting, the latter was considered to be a skill the teachers "already had attained". Such statements were shared with us during the meetings and written

down on the evaluation forms we collected after each meeting. We addressed teachers' assertions on these practices (acknowledging teachers' familiarity with Direct Instruction) but underlined the importance of implementing this effective instructional practice and emphasized that the implementation of the latter part of this model (i.e., at the end of a lesson) could be improved in the lessons of many participants in our program. A similar trend applied to the implementation of differentiation. Regardless of this attention for teachers' familiarity and experience and the constructive feedback we provided to individual teachers, this appeared to be insufficient to facilitate change in behavior. Perhaps participants were not convinced of the importance of these instructional practices (a topic addressed in, e.g., Deci, 2009), or they were under the impression that they mastered the skill at hand despite our feedback (not critically reflecting on their own implementation, e.g., Gay & Kirkland, 2003). Or perhaps other causes, such as the "rigidity" of the Direct Instruction model (mentioned in Baumann, 1988) or low levels of self-efficacy in relation to struggling students – impeding differentiation - (suggested in Paine, 1990 in Tomlinson et al., 2003) might be the reason for this limited implementation. Such aspects that impede implementation must be identified; this will help future PD programs that target the application of Direct Instruction and differentiation, and will add to the body of knowledge on teacher change and the instructional practices at hand.

Certain limitations with respect to our research design should be considered. In this chapter, we focused on teacher implementation data but we studied only a small part of an extensive multicomponent PD program. We considered the instructional practices of Direct Instruction, modeling, and differentiation to be the most likely aspects of teachers' instruction to have changed as a result of following the program: we actively targeted these practices during the meetings and they were expected to be fostered by other elements of the program. Yet PD programs aimed to change behavior might change knowledge and skills, but leave the actual behavior unaffected. Desimone (2009) proposed a conceptual model on evaluating PD programs in which teachers' knowledge, skills, and attitude should be measured prior to assessing classroom instruction. Doing so would have provided a more complete picture on the impact of the entire PD program on teachers, and would have more clearly identified which aspects of the program did work as expected and which aspects did not. Extending the research design to focus on more aspects than only Direct Instruction, modeling, and differentiation would have resulted in a more complete picture on teacher change as a result of the program.

A second limitation in our research design we would like to acknowledge is the fact that no observations were conducted in the control condition. Hence we cannot ensure that any difference between pre- and postmeasurement is completely attributable to the PD program. Alternative explanations (such as history or maturation, Cozby, 2003) might apply to the finding of increased implementation of modeling as identified in our study.

Despite these limitations, valuable information on implementation was collected. Not many teachers improved in their implementation of Direct Instruction, modeling, and differentiation, but teachers who did implement modeling were found to have higher students' results. We would like to end this paper by underlining the importance of studying teacher change. In the words of Black and Wiliam (1998b), in their paper on the "black box" between educational reform and student outcomes: "[h]ow can anyone be sure that a particular set of new inputs will produce better outputs if we don't at least study what happens inside?" (p. 140). In this chapter, we explored several aspects of how we assumed the multicomponent PD program would foster student reading comprehension and collected valuable information with respect to our hypothesized theoretical framework.

6. General conclusion and discussion

6.1 Introduction

The studies presented in this dissertation were conducted to evaluate the effectiveness of a multicomponent teacher Professional Development (PD) program. The teacher PD program aimed to improve students' reading comprehension in second and third grade. It was developed following recent concerns on Dutch students' reading comprehension performance (Ministry of Education, 2008; 2010) and the widely established importance of early proficiency in reading (e.g., Kirsch, 2002; Snow et al., 1998). The performances of Dutch students' on both national and international reading assessments were considered insufficient, and these unsatisfactory results had been ascribed to various causes namely the lack of clear performance goals for teachers and schools to aim for in their teaching (Council of Education, 2007; Inspectorate of Education, 2011; Ministry of Education, 2010), the quality of teachers' reading comprehension instruction which could be improved in terms of explicitness (Aarnoutse & Weterings, 1995; de Jager et al., 2002; Van Elsäcker, 2002) and differentiation (Inspectorate of Education, 2012; Van Berkel et al., 2007; Van Elsäcker, 2002), and the hypothesized difficulty of teaching reading comprehension due to the complexity of reading comprehension skills and inadequate curricular textbooks (Droop et al., 2012; Houtveen & Van de Grift, 2012; Stoeldraijer & Forrer, 2012).

The teacher PD program was developed to support teachers in changing their existing routines and help implement more effective instructional techniques. The rationale behind the program was that students' reading comprehension was expected to improve by making teachers' instruction more goal-oriented, focused, clear, and better suited to students' needs. For this purpose, the PD program was designed to contain three components namely, 1) setting standards and performance goals for every student, 2) applying formative assessment and data use, and 3) acquiring relevant instructional skills and (content and curriculum) knowledge in reading comprehension. Each of these components had separately been shown to be positively related to student performance (for working with goals, see e.g., Fuchs, Fuchs & Deno, 1985; for formative assessment and data use, see e.g., Carlson et al., 2011; for explicit instruction in reading comprehension, see e.g., Andreassen & Braten, 2011). We integrated these three components into one synergetic package, as the components were assumed to foster the desired instruction in an inter-related manner: particularly the implementation of Direct Instruction, modeling, and differentiation were assumed to be stimulated in this way. In the design of the program, the findings of studies on effective professional development were incorporated by, among other things, targeting collective participation of staff members from the same school and by stimulating active learning as teachers were provided with constructive feedback after lesson observations.

The multicomponent teacher PD program targeted second- and third-grade teachers from the same school as well as the school's principal and internal support coordinator. A pilot study was conducted in the school year of 2010-2011 which helped to refine the program's design and materials. The main study was conducted in the school year of 2011-2012, in which nineteen schools in the northern part of the Netherlands participated. In total, 33 second- and third-grade teachers (teaching 451 students) participated, and the school principals and internal support coordinators of these nineteen schools took part in the program as well. The time investment for the PD program was scheduled for approximately 40 hours, including attending nine after-schools meetings and completing the associated homework assignments.

In this dissertation, we addressed the following research question to evaluate the teacher PD program's effectiveness: *Does students' reading comprehension improve after teachers have followed a multicomponent professional development program targeting goals, data use, and instruction, and can we find further empirical evidence for the assumptions underlying this program?*

To answer this research question, we investigated the effects of the program on students' reading comprehension (Chapter 2). Furthermore, we focused on the teacher-set performance goals which played an important role throughout the PD program. Specifically, we studied the topic of validity with respect to the performance categories on which the goals were based (Chapter 3) and we investigated the relation between the performance goals and students' reading comprehension results (Chapter 4). Last, we focused on teachers' implementation of the instructional practices of Direct Instruction, modeling, and differentiation which were targeted in the PD program, and linked this implementation data to students' performance (Chapter 5). A summary of the main findings is provided next.

6.2 Summary of main research findings

6.2.1. The effect of the PD program on students' reading comprehension

In Chapter 2, we concentrated on the effect of teachers' participation in the PD program on students' reading comprehension. The propensity score matching approach (e.g., Rosenbaum & Rubin, 1985) was used to construct an equivalent control condition from a larger pool of possible controls. It was found that students in the experimental condition performed significantly better than those in the control condition on the Cito standardized reading comprehension assessment, with an effect size of $d = .37$, 90% CI [$d = .20$; $d = .55$]. We checked for the robustness of these results using different model specifications, and found similar albeit smaller effect sizes for the effect of the PD program on student achievement ($d = .29$, $d = .30$, and $d = .31$, respectively). According to Cohen's interpretation (1988), these effect sizes can be interpreted as small to medium effects. Differential effects of the program on student achievement were investigated but

these were non-significant: all students, irrespective of whether they were initially low or high achieving readers or whether they were in second or third grade, appeared to have profited equally from their teachers' participation in the PD program.

6.2.2. Investigating the performance goals

After identifying the effect of the program on students' results, we focused on the teacher-set performance goals as these goals were referred to and re-examined throughout the PD program; helping the teachers to attain their goals was an important rationale for the program's second and third component. At the end of the program, the teachers received an overview of the degree to which they had attained their own goals.

In Chapter 3, we focused on the performance categories on which the performance goals were based. With help of a standard setting procedure which entailed various rounds, the participants were asked to identify the four cutscores that marked the boundaries between the five successive categories (labeled below minimum, minimum, basic, proficient, and advanced: in acknowledgement of differences in students' capabilities). For our PD program, we were interested to what extent the cutscores were considered to be accurate – this was an assumption of the program which we wanted to investigate. To help evaluate the accuracy of cutscores, evaluation guidelines recommend investigating the evidence for different types of validity (Cizek & Bunch, 2006; Hambleton & Pitoniak, 2006; Kane et al., 1999; Norcini & Shea, 1997; Pant et al., 2009). We followed these guidelines and assessed the cutscores' procedural validity and internal validity. The procedural validity was studied with help of participants' feedback pertaining to a) the procedure's explicitness, b) the procedure's practicability, and c) the panelists' own deliberateness for setting cutscores. The internal validity was assessed through the investigation of the variation in cutscores across different rounds of the standard setting procedure; here, we studied d) the panelists' adaptations across rounds, e) the correspondence between cutscores and empirical performance data, and f) the interpanelists' agreement. The results of our analyses indicated that both types of validity were supported.

In Chapter 4, we focused on the teacher-set performance goals. In order to assist teachers in their task of setting goals for each individual student in their class, a multistep procedure was developed. This procedure aimed to help teachers reflect on and reconsider the goals' appropriateness with help of data use and team discussion before deciding on the final version of the performance goal. We assessed the use of this procedure by evaluating whether the final goals were equal to the goals they had set initially (i.e., at the beginning of this procedure) or whether these final goals were different. In the evaluation of the multistep procedure, a significant amount of change between the initial and the final goals was found. This result was considered to be indicative of the deliberateness of the final goals. After this, we evaluated the relation between the performance goals and students' achievement - this relation was presumed in our program,

and we wanted to investigate this. For this purpose, we assessed the degree to which the goals were attained by the students and whether the goals were significant predictors of student achievement while controlling for relevant student and classroom level covariates. The performance goals were attained by 79 percent of the students, as they performed at the desired level or higher. In addition, the performance goals were found to be significant predictors of performance. Higher goals were associated with higher results, a finding which concurred with the literature on goal setting (e.g., Locke & Latham, 1990; 2002) and teacher expectancies (Jussim & Harber, 2005; Rosenthal & Jacobson, 1968). This positive effect of high goals on achievement was found to be even stronger for initially low-achieving students, another finding concurrent with the literature (Good & Brophy, 2003).

6.2.3. Exploring teachers' implementation of instructional practices

In Chapter 5, we focused on teachers' implementation of specific instructional practices which were targeted in the multicomponent PD program. To help make reading comprehension instruction more goal-oriented, focused, clear, and better suited to students' needs, we trained teachers in the use of a) Direct Instruction, which is a teacher-centered model for instruction focused on the content and structure of a lesson, b) modeling, which is an instructional technique in which the teacher demonstrates how to solve a problem or apply a reading strategy by thinking aloud, and c) differentiation, which is an instructional practice in which the teacher attends to differences in student needs via the provision of extended instruction and by differentiating in assignments. We wanted to investigate the assumption that the implementation of these instructional behaviors would improve after following the program. Comparing observation results collected in the fall of 2011 to those collected in the summer of 2012, a significant improvement was found for the number of teachers who implemented modeling. Furthermore, students whose teachers modeled showed significant higher assessment results than students whose teachers did not model with $d = .24$, 90% CI [$d = .03$; $d = .46$], although it must be acknowledged that the number of teachers who implemented modeling on the postmeasurement was relatively small. The implementation of both Direct Instruction and differentiation was found to be rather limited and did not change from pre- to postmeasurement.

6.3 Discussion

6.3.1. Limitations

In this section, we would like to acknowledge certain limitations which were pertinent to the overall study as discussed in this dissertation. First of all, in the dissertation it was concluded that the teacher PD program was effective in improving students' achievement but the question *how* the program succeeded in improving student achievement has been left mostly unresolved. Aiming to improve students' reading comprehension, we set out to make teachers' instruction

more goal-oriented, focused, clear, and better suited to students' needs. This was targeted through the use of the three components and, in our view, represented in teachers' implementation of Direct Instruction, modeling, and differentiation. By stimulating the implementation of Direct Instruction (particularly the elements of this model pertaining to the beginning and end of a lesson), we assumed to make instruction more focused. By stimulating the implementation of modeling, we expected instruction to become clearer. And by stimulating the implementation of differentiation (in terms of extended instruction and differentiation in assignments) in relation to the student-specific performance goals, we expected teaching to become more goal-oriented and better suited to students' needs. At the end of the PD program, teachers' implementation of these three instructional practices was found to be limited. As a result, we cannot attribute the higher reading results of students taught by teachers in the experimental condition to these teachers' implementation of the targeted instructional practices. The use of a more elaborate observation instrument, assessing more aspects of instructional quality and distinguishing between different levels of quality (rather than dichotomous questions on occurrence) would have been favorable, particularly as we targeted more aspects throughout our program than only Direct Instruction, modeling, and differentiation. Nonetheless, a positive effect of the program on students' reading was demonstrated for which we assume teachers' instruction to have benefitted from the program as students' reading comprehension is not easily improved. Reading comprehension is known to be a complex active and interactive process (e.g., Afflerbach et al., 2008; Snow et al., 1998) and not all teacher professional development programs targeting reading succeed in significantly improving student results (see the overview of studies in Yoon et al., 2007). In short, we expect the multicomponent PD program to have improved the quality of teachers' instruction on aspects which we did not measure with our observation instrument. In the paragraph on suggestions for future research, recommendations are provided concerning the identification of those elements of the program which are expected to have positively influenced students' reading comprehension.

One might consider the *Hawthorne effect* (i.e., participants improving their behavior simply because of the knowledge that they are being studied, and not because of the content of the program; Shadish et al., 2002) as an alternative explanation of the positive effects on students' reading comprehension. Yet as previously addressed in Chapter 2, we consider this less probable. The positive effect of the program on student achievement was identified using the propensity score matching approach in which we selected an equivalent control group from a larger pool of possible controls. All schools in the conglomerate of intervention studies participated because they wanted to improve their education, and all schools and teachers were aware of students' results being measured throughout the entire school. Despite the fact that school and teacher characteristics could not be taken into account in the construction of the propensity score (due to non-response on a questionnaire), we are of the opinion that relatively similar schools and teachers participated in both the experimental and the control condition. Nevertheless, a

replication of this study using random assignment to conditions would complement the findings of the current dissertation as stronger statements on the causation of the program's effects could then be made (in that case, there would be no threat of omitted variable bias with respect to the participating teachers and schools). In such a replication, the investigation of retention effects on students would also be recommended as the current research design did not include follow-up measures.

The last limitation addressed here pertains to the fact that the researchers who conducted the evaluation on the effectiveness were the developers of the program and the facilitators during the meetings. The advantage of this design (researchers being the facilitators) was that data collected during observations was used as input for the meetings and part of the coaching-aspect of the program. In light of possible *experimenter bias* (e.g., Rosenthal & Fode, 1963) it would have been recommended that other researchers would have evaluated the effectiveness of the program, to better ensure objectivity in relation to the current findings. Yet studies that evaluate the effectiveness of newly developed programs must first focus on what Borko (2004) calls "an existence proof" (p. 5), by investigating the evidence that a professional development program can have a positive impact on relevant outcomes. Only after this existence proof is it considered worthwhile to have external researchers conduct evaluations on the effectiveness of the program, as the program is then presumably implemented on a larger scale. In such a case, we would recommend the involvement of external researchers.

6.3.2. Directions for future research and implications for teacher training

Taking into account the current findings as well as the limitations of the different studies presented in this dissertation, several directions for future research are proposed. With these recommendations, we focus on making the program more effective and more efficient. After this, several implications for teacher PD programs and teacher training are addressed.

6.3.2.1 Making the program more effective and efficient by focusing on teacher implementation and change in instruction

First of all, taking sufficient time to ensure teacher implementation prior to measuring effects on students is recommended. In the evaluation literature, this is referred to as the *fidelity of the program's implementation*, and more specifically, to participants' *adherence* to deliver the content of the intervention as designed (O'Donnell, 2008). Other programs aiming at reading improvement, such as the Improvement of Reading Comprehension Quality (Houtveen, 2002) or Success for All (Slavin, 2002), deliberately take out multiple years to ensure sufficient implementation on the side of the teachers before evaluating the effectiveness of the program. In this way, the effectiveness of the program pertains to the effectiveness of teachers' *instruction as presumed by the program*. The effect of the multicomponent PD program on achievement as

identified in Chapter 2 is considered remarkable given the fact that presumed instructional practices were implemented to a rather limited extent. Larger effects of the program are expected if the implementation of modeling, Direct Instruction, and differentiation would have been stronger. Training teachers during one or two school years and measuring effects on students in the following school year (including necessary follow-up meetings and observations to ensure sufficient implementation in this last year) would have given teachers more time to change their existing routines; a process which is known to be difficult (e.g., Fullan, 2001). Simultaneously, we expect some modifications to the program to be necessary before such teacher change can be expected. For instance, it might be required that teachers receive feedback on their instructional behavior more frequently than currently done during the PD program. Modifications of the curricular materials might be required as well. Moreover, specifically for the implementation of Direct Instruction and differentiation, teachers' attitudes and beliefs were suspected to have played a role in hindering the implementation of these practices; therefore, during the preparation phase of the program and at the start of the school year(s) in which teachers are trained, perhaps a more explorative approach is suitable in order to better identify (possible) implementation problems so that they can be tackled later on in the program. When following these recommendations - thus, conducting the program again and focusing on teachers' implementation of Direct Instruction, modeling, and differentiation first -, we propose to simultaneously focus on identifying which other aspects of instruction are affected by the PD program. For example, during the PD program we focused on key concepts in second- and third-grade reading comprehension and focused on using performance data to guide instruction: it might be that these aspects improved the quality of instruction causing the students in the experimental condition to outperform the students in the control condition (for instance, by having lessons that are more rich in content or in which students receive more targeted questions during instruction). Using a mixed methods approach for a selection of teachers and their classes (in which, for instance, interviews, analyses of group plans, and video-observations are conducted) will help to identify whether these aspects have resulted in higher levels of instructional quality. If this appears to be the case, these aspects can then be addressed in the PD program in a more targeted way. This will improve the efficiency of the program. More recommendations on how to make the program more efficient are provided further on.

6.3.2.2 Making the program more effective for different groups of students

A second recommendation would be to develop and evaluate modifications of the PD program which are aimed at the performance improvement of specific subpopulations of students, namely the weaker performing students and the well achieving students. Given the fact that this program was developed out of concern on the degree to which struggling, poorly performing readers were prepared for later schooling (Inspectorate of Education, 2007; 2010b), it is considered worthwhile to adapt this program in such a way that the performance improvement of

this subgroup is facilitated and sufficient reading skills can be guaranteed. Students who initially scored weak on reading comprehension appeared to benefitted extra from higher goals, but their teachers might need more specialized support in order to set higher goals and attain them with these specific students. In contrast, high achieving students did not appear to benefit strongly from the use of performance goals in our study; a finding which was expected to be caused by a ceiling effect. Recently, the relative underperformance of strong readers and limited degree of challenging instruction and assignments these students receive (Doolaard & Harms, 2013; Meelissen et al., 2012) has become a topic of general concern (e.g., Inspectorate of Education, 2013). Adapting the program in such a way that the learning opportunities of both weak and well achieving students are fostered is considered to be a valuable endeavor.

6.3.2.3 Making the program more efficient by making it more adaptive to its participants

A third recommendation for future research would be to investigate whether the program can be better suited to the participating teachers' skills, knowledge, and attitudes. While facilitating the current PD program, we already made a few modifications in acknowledgement of differences between teachers and school teams. If we would assess teachers' skills, knowledge, and attitudes beforehand, we could provide school teams with a version of the program which is better suited at the school's participating teachers. This will improve the quality of these teachers' instruction and improve the efficiency of the degree to which these teachers better meet individual students' needs; hence, students' performance improvement is realized in a more efficient way. Perhaps the program can be modified in such a way that the school principal or internal support coordinator can facilitate the meetings, adapting the content of the program in such a way that it is better aligned with the school's own teacher and student population. For example, Success for All makes use of an on-site, full-time program facilitator who oversees the daily operation of the program and who coordinates several of the program's components (Slavin, Madden, Chambers, & Haxby, 2009). In order to realize this stronger alignment between the content of the program and the teachers' background, it is recommended to establish teachers' level of skills, knowledge, and attitudes prior to the start of the program. This can be done by having the school principal or internal support coordinator organize meetings to discuss the participating teachers' prior attained skills, knowledge, and beliefs regarding the aspects at hand.

6.3.2.4 Implications for teacher professional development and teacher training

The current program provided in-service training to teachers using a multicomponent PD program. In Chapter 2, we discussed the review study of Yoon et al. (2007) and the Success for All program (Slavin, Madden, Chambers, & Haxby, 2009) in which the most promising results of teacher professional development are found for programs that combine a subject-specific focus with data use. To further students' performance via teacher professional development, it is assumed that a combination of these components is required in order for the program to be

successful. Our program was an example of such a multicomponent approach. Researchers and practitioners in the field of teacher PD are recommended to move on along this path.

Furthermore, the integration of the content of our PD program into initial teacher training programs is considered a useful activity. Many aspects, such as working with performance goals, a formative approach to using data and a stronger knowledge base for reading comprehension instruction can be incorporated in teacher training programs. In addition, training in the use of the student monitoring systems as well as training in instructional practices such as Direct Instruction, modeling, and differentiation is considered feasible for this context as well. In this way, beginning teachers are equipped with important knowledge and skills for teaching reading comprehension.

6.3.2.5 Concluding note

This dissertation focused on the evaluation of a multicomponent teacher PD program aiming to improve Dutch students' reading comprehension in second and third grade. The results of the studies reported in this dissertation as well as similar studies targeting performance improvement are considered an important contribution to the field of educational science (also in Borko, 2004) as they "evoke images of the possible (...) not only documenting that it can be done, but also laying out at least one detailed example of how it was organized, developed, and pursued" (Shulman, 1983, p. 495).

Appendices

- Appendix 1: Overview of the professional development program and specifications per meeting
    - 1. Overview of the multicomponent PD program
    - 2. Scheduling of the PD program's meetings
    - 3. General set-up of each meeting
    - 4. Specifications of the content per meeting
- Appendix 2: The Dutch educational context in relation to the PD program's three components
    - 1. Recent implementation of standards in the Netherlands
    - 2. The Cito LOVS standardized assessment system and student monitoring systems in the Netherlands
    - 3. General performance expectations, common instructional practices, and the curriculum
- Appendix 3: Overview of the cutscores and performance categories
- Appendix 4: Description of the data analyses used in the PD program
- Appendix 5: Lessons learned from the pilot study
- Appendix 6: Evaluation form used during the standard setting meeting
- Appendix 7: Observation instrument

Appendix 1: Overview of the professional development program and specifications per meeting

The studies presented in this dissertation were conducted to evaluate the effectiveness of a multicomponent teacher Professional Development (PD) program. This PD program aimed to improve students' reading comprehension and students' mathematics due to performance concerns for both areas (Ministry of Education, 2009; 2010). All aspects of the program which have been reported in this dissertation in relation to reading comprehension (for example, setting standards and performance goals) have also been conducted for mathematics. The content of the Appendices will pertain to the PD program as it was delivered to the participants, and will thus pertain to both subject areas. The results of the program with respect to mathematics will be discussed further in the dissertation of Ritzema (forthcoming).

In this section, the reader can find detailed information on the overall rationale behind the PD program, its aims, and the way it has been conducted. This overview provides, for example, empirical support and practical arguments underlying the specifications of the program. In this way, we want to give a clear account of how our program was realized and provide other researchers with the necessary information to replicate our study. The following sections and paragraphs therefore mainly contain information of specific interest. First, we will provide an account of the entire PD program. Next, an overview of the program's meetings is presented, and several general characteristics of these meetings - such as delivery format and duration – will be discussed. After this, each meeting is dealt with separately, describing its aim, content, the materials used, and the related homework assignments.

1. Overview of the multicomponent PD program

The teachers that participated in the PD program were supported in improving their practice with help of a three-component program: 1) setting standards and performance goals for every student, 2) applying formative assessment and data use, and 3) acquiring relevant instructional skills and (content and curriculum) knowledge in reading comprehension and mathematics. All three components have shown to be positively related to student performance (see Chapter 2). The PD program was designed to foster student learning through teachers' application of a multicomponent package, which aimed to make instruction more goal-oriented, focused, clear, and better suited to students' needs. The three components were integrated into one synergetic package, as the components were assumed to foster the desired instruction in an inter-related manner. In the paragraphs below, the information on each of the three components is briefly resumed (for more information, see, e.g. Chapters 1 and 2 in this dissertation).

1.1. Component 1: Setting standards and performance goals for every student

Goal setting was incorporated as the first component in the PD program as the insufficient results of Dutch students on both international and national assessments were attributed to the fact that, for schools and teachers, it was unclear what students should know and do at certain time points (Expert group Continuous Learning Progression, 2008). Setting goals leads to a clearer notion of how success can be attained and it focuses the attention on the realization of relevant outcomes (e.g., Fuchs et al., 1985; Locke & Latham, 1990). The goals were based on performance categories which had been identified by the participants with help of a *standard setting procedure* (see Chapter 3). As discussed in Chapter 4, we aimed to assist teachers in setting goals that were 'difficult but not too difficult' given their students' capabilities - referring to the prior reported results of Erez & Zidon (1984, in Locke & Latham, 1990). For this purpose, we developed a *multistep procedure* which incorporated performance data analysis and team discussion to help teachers reflect on and reconsider the goals' appropriateness before deciding on its final version - following recommendations of the data use literature (e.g., Schildkamp & Kuiper, 2010). These aspects pertain to the second component of our program.

1.2. Component 2: Applying formative assessment and data use

In order to help the participating teachers set and attain the performance goals, it was important that they based their instructional decisions on assessment results (e.g., Guskey, 2002). Using student performance data to adapt one's teaching in order to better meet students' needs is known as *formative assessment* (Black & Wiliam, 1998a; 1998b; Herman et al., 2010). The participants therefore received training in the use of the student monitoring system. Yet the concept of *performance data* not only pertains to the assessment results on standardized tests, but also, for example, to completed work book assignments or teacher observations of how the students function in class (Lai & Schildkamp, 2013). Teachers' reflection was targeted by focusing their attention on different sources of data they could use which would help to better suit their instruction to different students' needs. Thus, by working with student-specific performance goals and monitoring performance in relation to these goals (i.e., components one and two), it was expected that teachers would attend more to different student needs and that differentiation would be fostered as a result. An important prerequisite, however, is that teachers adjust their practices accordingly after analyzing the data, a step which is not always guaranteed (Goertz, Olah & Riggan, 2009 in Carlson et al., 2011). The third component of our PD program focused on how to take action after analyzing the data.

1.3. Component 3: Knowledge, instruction and the curriculum for reading comprehension

In the PD program, after setting the performance goals and identifying the progress made toward them, it was important to help the teachers attain their own objectives by ensuring that they were sufficiently equipped with the most relevant instructional skills and knowledge about

reading comprehension and mathematics. We targeted *Direct Instruction* (a teacher-centered model of instruction) and *modeling* (an instructional technique in which the teacher demonstrates how to apply a certain reading comprehension or mathematics strategy by "thinking aloud"; to show students which strategies are appropriate, how to pursue them and why). After having been introduced to these concepts, the teachers practiced and received feedback on the implementation of the instructional approaches. In addition, the participants were informed on the set-up of the Cito-assessments for mathematics and reading comprehension in the grades under study. Here the degree of alignment between these assessments and the curricular text books used in their schools was discussed and tips were provided on how to bridge evident gaps. Specifically for reading comprehension, we discussed important determinants of reading performance and key concepts in the second- and third-grade reading comprehension curriculum.

In Figure 1, the interrelatedness of the components is illustrated. This graphical representation is given at the beginning of the detailed descriptions of the meetings (presented below). For each meeting, the most essential component is highlighted.

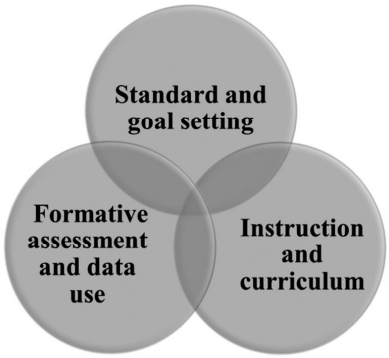


Figure 1. The PD program's three components

2. Scheduling of the PD program's meetings

Throughout the school year, the time investment of the teachers was scheduled for 40 hours, including attending the nine after-school meetings and homework assignments. In Figure 2, a graphical overview of the nine meetings is presented. The three components of the PD program were addressed to comparatively the same extent. The standards and goal setting component (the first component) was addressed in meetings 1, 3 and 4, respectively. Information and training on formative assessment and data use (our second component) was dealt with in meetings 2, 5, and 9.

Training in relevant instructional practices and information to improve teachers' (content and curriculum) knowledge (the program's third component) was targeted in meetings 6, 7, and 8. In the majority of the nine meetings, the subject areas of mathematics and reading comprehension were targeted simultaneously. In meeting 1 and 6, however, the emphasis was specifically on mathematics, while meeting 3 and 7 particularly addressed reading comprehension.

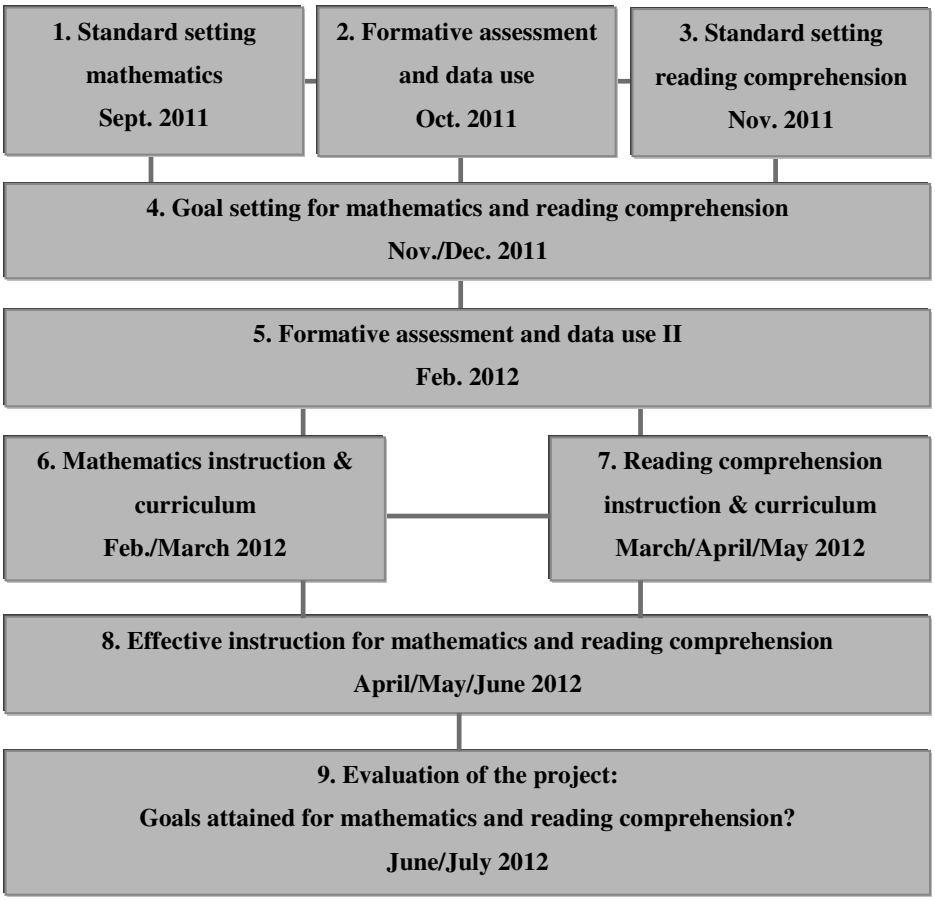


Figure 2. Overview of the meetings

3. General set-up of each meeting

The standard setting meetings and the meetings on formative assessment and data use (meeting 1, 2, 3, and 5) were set up as general gatherings which were held in a convention centre. These meetings were scheduled to last 2.5 hours. The other meetings (meeting 4, and 6 to 9) took place at the individual schools; in a few cases, the participants of two or three different schools joined together in one meeting. These meetings were scheduled to last 1.5 hours.

All meetings followed the same order: a) recapitulation of the last meeting, b) overview of the content to be discussed in the current meeting, c) presentation of information, d) recapitulation of the current meeting, and e) preview of the next meeting. The structure of these meetings resembled the elements in the Direct Instruction model, an effective instructional approach which was also discussed during the meetings (also detailed in Chapter 5).

Different delivery modes were used during the meetings. We provided short lectures using whole-group (power point) presentations. During several of these whole-group presentations, video-fragments were shown as illustrative material. In addition, during almost all meetings, the participants were asked to work on assignments "on the spot". These assignments could be individual assignments, assignments requiring collaboration between colleagues from the same school, and assignments requiring collaboration between colleagues from other schools. Use of the student monitoring system (part of our meetings on data use) was practiced behind laptops. For the entire program, the majority of the hand-outs and other materials offered were self-developed. To organize the hand-outs and materials which were distributed during the meetings, we provided a binder to all participants.

At the beginning of each meeting, the attendance of the participating teachers, school principals and internal support coordinators was registered. At the end of each meeting, the participants were asked to fill in an evaluation form about how they had experienced the value and practicality of that particular meeting. During the standard setting meetings (meetings 1 and 3), specific questions were posed about the different rounds in the standard setting procedure and the degree to which the participants considered their own cutscores as well-considered. This was done because on-site evaluations by participants serve as an important check on the validity of the cutscores and the way in which they have been set (Cizek & Bunch, 2006; Hambleton & Pitoniak, 2006) – also see Chapter 3.

4. Specifications of the content per meeting

In this section, the content of the nine meetings will be addressed separately. For each meeting the most vital component is highlighted in the figure next to the headings.

1. Introduction to the project and standard setting for mathematics

Summary of the first meeting: The goal of this meeting was a) to inform the participants on the set up of the PD, and b) to set cutscores and create performance categories for the second and third grade (end-of-the-school year) June-mathematics assessments using a standard setting procedure. This procedure was also conducted to facilitate and stimulate the teachers' awareness of their own performance expectations and the instructional and curricular demands for the second- and third-grade mathematics.

Characteristics	
Format of meeting	General meeting
Time span	2.5 hours
Delivery mode	Whole-group presentation, small-group discussion
Hand-outs	Ordered Item Booklet, standard setting forms, training materials for standard setting procedure, print-outs of powerpoint slides

1.1 Introduction to the project

At the beginning of the first meeting, a short outline of the project was presented to the participants. We briefly introduced the three components of the PD program as well as their timing throughout the school year. In addition, the participants were informed about certain data collection obligations related to partaking in the PD study (i.e., , having a pre- and post-observation during the mathematics and reading comprehension lessons, filling in a questionnaire at the beginning and at the end of the PD program, and providing performance data to the researchers prior to several meetings).

1.2 Introducing performance standards

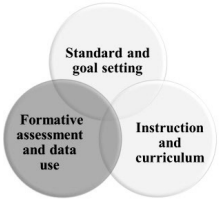
The participants were explained that by specifying what students should be able to know and do, teaching would be directed toward the attainment of desired outcomes. We argued that having clear performance goals makes it easier to target instruction toward the attainment of these objectives (Fuchs, Fuchs, & Deno, 1985) and that this targeted instruction is expected to improve student results (Lauer et al., 2005; Roeber, 1999)[28]. For this purpose, the participants were asked to participate in the standard setting procedure, in order to establish performance categories. The benefit of working with goals that are based on performance categories was that the attainment of these goals would be easily established by conducting standardized assessments in class. Furthermore, we provided information on the current Dutch educational policy regarding the performance standards (as discussed in Appendix 2). By introducing them to this background, we hoped that the participants would recognize the value of working with performance goals, which would in turn positively influence their commitment to the PD program.

1.3 Standard setting for mathematics

Within the PD program, the Cito adaptation of the Bookmark procedure (see Van der Schoot, 2009) was used. Here, participants considered a selection of items from the Cito mathematics assessments and indicated which items they expected the students to answer correctly in the June-assessment: for the remainder of this Appendix, the end-of-the-school year assessment is referred to as the *June-assessment* and the midway-of-the-school year assessment is referred to as the *January-assessment*. To facilitate the standard setting task, the items were ordered in such a way that they increased in difficulty and we presented them in a so-called *Ordered Item Booklet* (OIB)[29]. The explication of performance expectations is done for students of different ability levels. In this PD, five performance categories were distinguished: below minimum, minimum, basic, proficient, and advanced. For each category, the participants decided at which item there was a suitable cutoff (between *below minimum* and *minimum*, between *minimum* and *basic*, etcetera). Thus, four cutoff points needed to be identified in total[30]. Since all items can be converted into scores on the assessment scale, indicating a pupil's level of proficiency, the cutoff points can also be converted into scores: these points are thus also referred to as *cutoff scores* or *cutscores*. The final cutscores were determined after three rounds. In the first round, the panelists individually studied the OIB and formulated cutoff scores based on their own opinion. In the second round, they came together in small groups (consisting of three to five people) and discussed their cutoff scores. Groups could reach consensus, but they did not have to. After this

[28] Positive findings are provided by the goal setting theory (Locke & Latham, 1990; 2002). For more information, see Chapter 4 of this dissertation.
[29] Both the items and information on their difficulty (for which Item Response Theory was used) were provided by the Netherlands Institute for Educational Measurement, with whom we collaborated in this study.
[30] Similar to the work of Roeleveld and Béguin (2009), four cutoff scores were set in the PD program.

small group discussion each participant reset his/her scores. Based on these reset scores, the median cutscores were calculated 'on the spot', while the participants listened to a further explanation of the PD project and the three components. In the third round, the average cutscores of the group were presented and compared to the actual performance data of the student population (their "empirical equivalents", see Chapter 3). These empirical data indicated the participants how realistic and ambitious their own cutoff scores were at that stage in the procedure. After this display, the panelists again reset their cutoff scores, after which the final ones were calculated by taking the median of the scores of the third round. This was done, however, at a later moment: the final scores were presented during the second PD meeting. After collecting the forms containing the final cutscores, the first meeting was brought to an end. More information on the standard setting procedure is provided in Deunk, van Kuijk, and Bosker (in press) and in Chapter 3 of this dissertation.



2. Formative assessment and data use

Summary of the second meeting: The goal of this meeting was to present the concept of formative assessment and data use. The participants also practiced different types of data analyses using the Cito student monitoring system (detailed further in Appendices 2 and 4).

Characteristics	
Type of meeting	General meeting
Time span	2.5 hours
Delivery mode	Whole-group presentation, practice of data analysis behind laptops (in pairs)
Hand-outs	Data analysis booklets (containing navigational directions and screen shots of the Student monitoring system), print-outs of power point slides, homework assignments for the teachers and the school principal/internal support coordinator

2.1 Recapitulation of meeting 1 and presentation of the final cutscores

At the beginning of the second meeting, the final mathematics cutscores were presented (their exact scores are presented in Appendix 3). The student performance at the level of the cutscores

was illustrated using several exemplary items. All performance categories (below minimum, minimum, basic, proficient, and advanced) and their accompanying test score intervals (also in Appendix 3) were discussed in order to explain to the participants how students' June-assessment results would be converted into one of the five performance categories.

2.2 A model for data use

Next, the standards were coupled to the other two components of the PD program and these three components were briefly re-discussed. The second component focused on the use of assessment results to improve instruction. The teachers were stimulated to use assessment data as feedback to help them modify their teaching activities in such a way that they met the individual students' needs and thereby furthered their development. This practice is also known as *formative assessment* (Black & Wiliam, 1998b). During the PD, the *evaluative cycle for data-driven teaching*, a model from Ledoux, Blok and Boogaard (2009) was used. This model was similar to models of formative assessment (Black & Wiliam, 2009; Carlson et al., 2011; Herman et al., 2010) and the well-known Plan-Do-Check-Act cycle (Deming, 1986). The evaluative cycle for data-driven teaching was used to help the teachers understand the concept of formative assessment and data use (our second component), and included goal setting (our first component) as well as the role of instruction (our third component). Teachers' reflection was targeted in this model as well, by focusing their attention on their students' performance in relation to their own teaching practice. The model consisted of five questions being:

1) What do I want to accomplish with my students?

2) Which sources of information can I use to map out the performances of my students?

3) How are my students performing based on these different sources of information?

4) What do these performance results mean? Can I interpret them?

5) How have I been teaching my students and does my approach and/or goals need to be adapted?

In order to "get acquainted" with this model, an introductory assignment was provided. The participants had to reconstruct the logic of the model by putting its elements (i.e., the five questions) in the right order. This activity was conducted in small groups consisting of participants from different schools. After the introductory assignment, the definition of "data-driven teaching" (a common translation of the Dutch term "opbrengstgericht werken") was discussed. Empirical evidence for employing a data-driven way of teaching (e.g., Carlson et al., 2011) was presented as well. The experiences of several Dutch school teams who already worked in a data-driven manner were illustrated using video-material (Primary Education Council, 2009). Next, we elaborated on each of the five questions of the evaluative cycle for data-driven teaching. With respect to the second question (Which sources of information can I use to map out the

performances of my students?), we provided additional information on the difference between the Cito-assessments and the assessments of the curricular textbooks - also see Appendix 2 for more information on this topic. With respect to the third question (How are my students performing based on these different sources of information?), we provided additional information on the differences between the various student monitoring systems (discussed in Appendix 2).

2.3 Making use of the student monitoring system

In the Netherlands, the use of student monitoring systems by teachers is still rather limited for analyzing problems and adapting instruction, and teachers who do use the student monitoring systems are often unaware of the possibilities for more sophisticated analyses (Ledoux et al., 2009; Meijer & Ledoux, 2011; Schildkamp & Kuiper, 2010; van der Kleij & Eggen, 2013). During this meeting, the participants were informed on three specific types of analysis which could be performed using data from the Cito assessments called a) *the estimation of future performance*, b) *an overview of performance in the previous school year*, and c) *a progress report* (analysis of academic growth). For their exact content, see Appendix 4A, B, and C. These three analyses provided information on (estimates of) individual students' performance and would be used in homework assignments later on in the PD program.

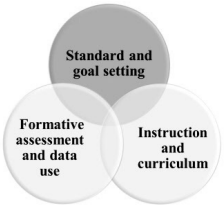
The participants practiced these analyses behind laptops (in pairs) while the researchers walked around to answer questions and give immediate feedback. The hand-outs for these exercises contained screen shots and directions on how to navigate through the student monitoring system. These exercises also contained additional information and questions to help the participants interpret and critically reflect on the output provided by the system.

2.4 Homework assignment

The participating teachers as well as the school principals and internal support coordinators - the latter two referred to as the *school management staff* - received booklets with specific homework assignments. The teachers were provided with a three-staged assignment, which focused on their own students' performance in mathematics. This three-staged assignment has already been discussed as part of the multistep procedure in Chapter 4; here, its content is recapitulated. First, we asked the teachers to intuitively predict their students' future performance on the June-assessment using the performance category classification; below minimum, minimum, basic, proficient, and advanced. Second, the participants were asked to use the student monitoring system to compute the future performance estimation, which is based on prior achievements (see Appendix 4A). Third, we requested them to use their student monitoring system to present an overview of last year's performance on the June-assessment (see Appendix 4B). Both the estimates and the performance scores could be converted to the performance category classification. The assignment booklet also contained questions about possible differences between the three assignments, like *'Are there differences between your own intuitive*

prediction and the Cito future performance estimate? If so, can you explain them?' These questions were developed to increase the teachers' affinity with the performance categories and to stimulate critical reflection on their own performance expectations. The homework assignment was requested to be completed prior to the program's fourth meeting, when the teachers would have to set performance goals for their pupils. Completion of the homework assignment would facilitate this task.

The school management staff was provided with a different assignment. The participants were asked to use the student monitoring system to compute the developmental growth (see Appendix 4C) of the current second- and third-grade students with respect to the prior school year, i.e. from the January-assessment to the June-assessment in grades 1 and 2. This assignment would offer more insight into how the students who were currently in second and third grades had developed the past school year. The assignment pertained to both the mathematics and the reading comprehension development of the students, thereby anticipating the subject area of interest in the next meeting. Similar to the homework assignment of the teachers, this task was requested to be completed before the fourth meeting.



3. Standard setting for reading comprehension

Summary of the third meeting: The goal of this session was to set cutscores and create performance categories for reading comprehension using the standard setting procedure, and to increase the participants' awareness of the performance expectations and instructional demands with respect to second- and third-grade reading comprehension.

Characteristics	
Type of meeting	General meeting
Time span	2.5 hours
Delivery mode	Whole-group presentation, small-group discussion
Hand-outs	Ordered Item Booklet, standard setting forms, training materials for standard setting procedure, print-outs of power point slides, homework assignments for teachers

3.1 Standard setting for reading comprehension

The cutscores for reading comprehension were set in the exact same way as those for mathematics. The only difference between both standard setting procedures pertained to the construction of the *Ordered Item Booklet* (OIB). The reading comprehension items increased in their difficulty, but the items that applied to the same text were grouped together to improve the booklet's readability. As reading comprehension concerns answering text-related questions, the difficulty of the question is influenced by (the difficulty of) the text. For example, the question "To what word does *this* in sentence 3 refer?" might be easy or difficult to answer depending on the complexity of the text. In Appendix 3, the final cutscores as well as the related performance categories are presented.

3.2 Reading comprehension development

In this part of the meeting, the complexity of reading comprehension was discussed. In comprehension processes, individual differences (e.g., decoding skills, prior knowledge) interact with the text features (e.g., text difficulty) (Ozuru, Dempsey, & McNamara, 2009; Snow, 2002). This interrelatedness between reader and text affects the ease with which one can concretize performance expectations in this domain. In order to aid the participants in determining what could be expected from second and third-grade students, the guidelines of the Expertise Centre for the Dutch Language (2010) were addressed (see Appendix 2 for their content). The Expertise Centre has identified seven skills that students in these grades are expected to master. All skills were briefly addressed, but the skill concerning students' genre knowledge was discussed more elaborately as it was part of teachers' homework assignment. The types of texts as discussed by the Expertise Centre for the Dutch Language (2010) were compared to those used in the Cito reading comprehension assessment. The types of questions used in these assessments were discussed as well, as some question formats were less common than those used in the curricular textbooks and workbooks (also in Appendix 2).

3.3 Homework assignment

The teachers were requested to identify (on a 3-point Likert scale) to what degree the text types and question formats in the Cito-assessment were similar to those in the curricular textbooks and workbooks as used in their classes. In the case of underexposure, the teachers were asked if they had ideas how to tackle this problem. We developed this homework assignment not to advocate teaching-to-the-test, but to improve teachers' awareness on these differences in order to acquaint the students with these question formats (so that students would not be unnecessarily surprised when taking the Cito assessments). The homework assignment was requested to be completed prior to the program's fourth meeting.

In addition, the teachers were asked to complete the same three-staged homework assignments for reading comprehension as was provided for mathematics. They had to 1)

intuitively predict their students' future performance on the June- assessment using the performance category classification, i.e., by selecting the below minimum, minimum, basic, proficient, or the advanced performance category, 2) use the student monitoring system to compute the future performance estimation - see Appendix 4A -, and 3) present an overview of last year's performance on the June-assessment - see Appendix 4B. This three-staged assignment was requested to be completed prior to the fourth meeting, as it was meant to facilitate the teachers in their goal setting process.

4. Goal setting for mathematics and reading comprehension

Summary of the fourth meeting: During this meeting the teachers had to set 'difficult but not too difficult' performance goals for each individual student in their classes, both for mathematics and reading comprehension. Furthermore, the teachers were stimulated to reflect critically on the curriculum alignment. Finally, for mathematics as well as for reading comprehension particular issues were addressed.

Characteristics	
Type of meeting	Meeting at school
Time span	1.5 hours
Delivery mode	Whole-group presentation, group discussion
Hand-outs	Print-outs of powerpoint slides, empty goal-setting table, print-outs containing specific examples of types of texts and question formats, print-outs containing exercises to practice the initiation of mathematics questions

4.1 Discussing the teacher expectations and performance estimates

During this meeting the homework assignments for mathematics and reading comprehension were discussed by the school team: the following information has already been discussed as part of the multistep procedure in Chapter 4. Here, we will recapitulate its content. The focus in this discussion was on explaining the differences between the teachers' original expectations of the students and the estimates from the student monitoring system. By discussing performance data, school team members can learn from each other and thereby improve their knowledge of and

skills in teaching (Datnow, Park, & Wohlstetter, 2007; Huffman & Kalnin, 2003; Schildkamp & Kuiper, 2010). For example, last year's teacher might ask the current teacher: *'Johnny was a rather good reader last year, and I would probably have expected him to perform on the proficient level. Why do you expect the basic level to be more suitable?'* By creating an open atmosphere in which the participants were willing to think along with one another, the supportive role of the school team was increased. Furthermore, a comparison was made between the expectations for students' performance in mathematics and in reading comprehension. The students' ability growth (part of the school management staff's homework assignment) was taken into account as well.

4.2 Setting performance goals

In this part of the meeting, the teachers were asked to set a performance goal for each individual student by reconsidering their initial expectations and taking into account all the information obtained during the group discussion. These performance goals were set for mathematics and for reading comprehension. They were formulated using the performance category classification (below minimum, minimum, basic, proficient, and advanced) and pertained to desired performance on the June-assessment. The teachers were instructed to set 'difficult but not too difficult' goals given the students' capabilities as these have been proven to be the most effective goals (Erez & Zidon, 1984, in Locke & Latham, 1990).

4.3 The school level: vertical curriculum alignment, school goals, and mutual expectations

After setting performance goals at the student level, we provided several recommendations, among which, to explicate which knowledge and skills are taught to the students in neighboring grades and how they are taught (Martone & Sireci, 2009; Webb, 1997). To illustrate this, we asked questions such as: *'Does the third-grade teacher know which skills are learned in grades two and four, and does he or she know in what ways these skills are taught to the students?'* Furthermore, we recommended the participants to openly discuss implicit expectations regarding instruction and the curriculum. These expectations could be either more general (*'In our school, we expect all our colleagues to have finished the curricular textbooks at the end of the year')* or content-specific (*'At the end of grade 2, we expect all students to master the multiplication tables 1-5')*. In addition, we encouraged the participants to think of content-specific goals that would complement the individual performance goals that were just set (e.g. *'At the end of grade 2, the basic/proficient and advanced students should be able to do automated additions up to 20'*). All these suggestions were meant to promote a continuous learning progression within the school.

4.4 Tips and practical suggestions on instruction

As part of this meeting, we discussed known difficulties in relation to the Cito tests and provided several practical recommendations.

(i) Mathematics: Word problems

As the curricular textbooks do not address how to solve word problems in a systematic manner (see Appendix 2), students tend to get confused when confronted with (complex) word problems in the Cito assessments (Janssen, van der Schoot, & Hemker, 2005). In the PD program, the teachers were instructed on how to initiate mathematical questions during other parts of the school day, for example, when students share personal anecdotes at the start of the school day. By initiating these questions, teachers can demonstrate to their students that mathematics is an omnipresent phenomenon and simultaneously demonstrate a systematic approach to solving such daily mathematical problems. Moreover, in this way the time spent on mathematics is extended. The participants were given an assignment in which they had to practice their ability to elicit mathematical thinking "on the spot". They were asked to read a short story about a Christmas tree and broken decorations, and come up with questions that targeted different mathematical domains (these domains are discussed in Appendix 2).

(ii) Reading comprehension: Text types and question formats

During this part of the meeting, the outcomes of the previous homework assignment (pertaining to the reading comprehension textbooks) were discussed. After a brief recapitulation of these text types and question formats, we discussed their (under)exposure on the basis of questions such as *'Do you have ideas on how to deal with certain underexposure?'* and *'Which suggestions do your colleagues have?'* Specific ideas about how to develop exercises that appeal to students' interests were mentioned. Furthermore, the teachers were supplied with a print-out with examples of different genres and question formats.

5. Formative assessment and data use

Summary of the fifth meeting: The goal of this meeting was to inform the participants on how to implement data-driven teaching at the levels of the school, the classroom and the lesson. Furthermore, they were given the opportunity to practice with different data analyses within the student monitoring system.

Appendix 1

Characteristics	
Type of meeting	General meeting
Time span	2.5 hours
Delivery mode	Whole group presentation, video fragments, small group discussions, practice of data analysis behind laptops (in pairs)
Hand-outs	Data analysis booklets (with navigational directions and screen shots of the student monitoring system), print-outs of powerpoint slides, print-out with nine propositions, homework assignment for the teachers

5.1 Data-driven teaching at the school, classroom and lesson levels: introduction

In our PD program, data use and the evaluative cycle of data-driven teaching entailed, among other things, setting performance goals for each individual student and making instructional decisions based on data so that the students obtain these objectives. Data-driven teaching is, however, not only limited to this classroom level. It also requires a change in school culture: important prerequisites for data-use include a clear vision of the school principal on issues such as the school's goals and teacher collaboration (Schildkamp & Kuiper, 2010; Wayman, Midgley, & Stringfield, 2006). Within the PD program, the need for school and teacher change was illustrated by showing video-fragments (Primary Education Council, 2009) of teachers, internal support coordinators and school principals who explained in which ways data-driven teaching had changed their way of working and their view on the school's culture. These fragments also illustrated *how* these people changed their way of working as a result of their increasing awareness of the students' performances and needs. After this general introduction, the participants split up into two groups: one groups with only teachers and one group with the school management staff (i.e., the school principal and internal support coordinator). The remainder of this meeting continued in parallel sessions which were conducted in separate rooms.

5.2 Data-driven teaching at the classroom and lesson levels: teacher session

The teacher session started with a short recapitulation of why data use is important. Several characteristics of effective schools were introduced, such as having high expectations in regard to student achievement and frequently monitoring and evaluating performance (Sammons et al., 1997; Scheerens & Bosker, 1997). Next, the teachers' reflection on their own data use and goal orientation was stimulated by having them conduct a small exercise. They had to score their own behavior by considering nine provoking propositions, among which "*When I look at the students' results I ask myself the following questions: do I see developments that I consider to be positive, do I see developments that I consider to be negative and are there any changes with respect to*

the previous assessment results?" After rating their own behavior, they were asked to discuss their own behavior in pairs. In order to elicit vivid and informative conversations these pairs were designed to consist of participants with different scores (high and low; colored cards represented these scores and these pairs could thus easily be constructed). In the next part of the meeting, we discussed the instructional model of *Direct Instruction*. This teacher-centered model was discussed as it has proven to be an effective instructional model (Borman et al., 2003). We explained the model's essential characteristics, while primarily focusing on the start and the end of the lesson (Leenders et al., 2010; Muijs & Reynolds, 2011). For more on this model, see Chapter 5. In addition, the topic of *time-on-task* was addressed. The amount of time teachers actually spend on *teaching* (instead of classroom managerial and organizational actions), and thus the actual time that students are *learning* determines student learning gains to an important extent (Houtveen, Van de Grift, & Creemers, 2004). We provided tips to help the teachers identify their own time-on-task and classroom management behaviors.

(i) Teachers' use of the student monitoring system

During this part of the teacher session, the participating teachers received training in making use of the student monitoring system (similar as in the second meeting of the PD program). Three types of analyses were discussed: a) *a performance comparison (to make extreme high or low scores more tangible)*, b) *identification of the average class performance for several consecutive years,* and c) *error analysis for mathematics* – analyses are described in detail in Appendix 4D, E, and F respectively. Similar to the second meeting, the participants practiced these analyses behind laptops in pairs with the help of hand-outs containing step-by-step instructions, explanations and critical questions about the outcomes. One of the members of the research team was available for assistance. As two schools had indicated beforehand that they were already quite familiar with the analyses which would be discussed during the current meeting, they followed a slightly altered program during this part of the meeting. They were exempted from several assignments but they were asked about their actual data use by a second member of the research team, who provided them with targeted information to improve their current data analyses practices[31].

5.3 Data-driven teaching at the school level: school management staff session

The school management staff session was set up as follows. For schools wanting to work in a data-driven way, the school management staff plays an important role. They are responsible for several tasks, such as monitoring the school outcomes, formulating clear goals, and promoting teacher collaboration (Schildkamp & Kuiper, 2010; Young, 2006). In the PD program, we stressed the importance of these activities. The structure of this meeting was similar to that of the

[31] Since the teachers indicated that they only worked with error analyses for a few individual students, further information on whole group error analysis was provided using the output of a fictitious group.

teacher session, as it started with the same recapitulation of characteristics of effective schools. Next the participants in this group were asked to respond to the nine propositions (comparable to the propositions the teachers received, but then formulated at the school level) on teachers' goal-orientation and collaboration. Beliefs and practices were discussed in pairs (again, consisting of participants with a high and a low score) to boost awareness and generate practical ideas on how data use could be implemented within the school. After this exchange of ideas, the implications for the school and the HR policies were discussed and illustrated by using video-material (Primary Education Council, 2009). It was further argued that the school management staff should play an important role in supporting the teachers' professional development (also in Fullan, 2001) by regularly observing classroom lessons and providing them with constructive feedback to improve their instructional practices.

(ii) School management staff's use of the student monitoring system

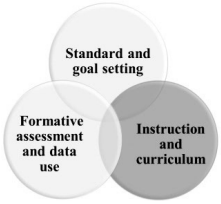
During this part of the meeting, the school management staff received training in working with the so-called self-evaluation module of the Cito student monitoring system. The three types of analyses discussed were a) *cross section of the performance levels of the middle school classes*, b) *longitudinal development of the group mean score of one or more classes,* and c) *the mean score of a specific grade* (e.g. grade three) *throughout consecutive years*. These analyses are elaborated in Appendix 4G, H, and I respectively. For the participants who were already more experienced in working with the student monitoring system, an additional type of analysis was offered, namely d) *group analysis* (elaborated in Appendix 4J), which examines the development of one class in terms of different subject areas and in several years. Again, the participants worked behind laptops in pairs, using hand-outs containing step-by-step instructions, explanations and critical questions about the outcomes. One of the members of the research team was available for questions and support.

5.4 Homework assignments

Both the teachers and the school management staff received a booklet with homework assignments. The teachers were given an assignment pertaining to the January-assessment for mathematics. They were asked to make two error analyses, one for the whole group and one for an individual student. The booklet contained instructions for executing the analyses in the student monitoring system. Guidelines on how to interpret the output were provided as well[32]. The booklet also included questions to stimulate one's reflection on the output, such as '*Can you indicate mathematical domains for which a vast amount of students scored below or above expectation? If so, can you explain why these deviations occurred?'* The assignment was requested to be completed prior to the sixth meeting.

[32] Since the output of an error analysis consists of two tables containing a vast amount of numbers, additional explanation was required on how to interpret this information properly.

The management staff was provided with a different assignment. For the second time in the PD program, they were asked to compute the developmental growth for the current second- and third-grade students with the help of their student monitoring system (the progress report, discussed in Appendix 4C). This time, they had to calculate the developmental growth from the June-assessment in the previous year until the current January-assessment. Similar to the previous progress report assignment, the developmental growth of both mathematics and reading comprehension had to be analyzed. The assignment was requested to be completed prior to the sixth meeting.



6. Mathematics instruction and curriculum

Summary of the sixth meeting: The goal of this meeting was to evaluate the January-assessment results for mathematics, to interpret the mathematics error analysis, to re-examine the teachers' performance goals for the June-assessment (and adjust them if desired) and to extend their knowledge of solving word problems in mathematics.

Characteristics	
Format of meeting	Meeting at school
Time span	1.5 hours
Delivery mode	Whole-group presentation, group discussion
Hand-outs	Overview of student achievements on January-assessment and performance estimates for June-assessment, overview of exercises regarding different domains, print-outs of powerpoint slides, homework assignment for teachers

6.1 Start of the meeting

At the start of this meeting, an overview was handed out to each class, consisting of the performance goals based on the five performance categories (below minimum, minimum, basic, proficient, and advanced). The aim was twofold: to provide a clear outline of the performance goals at the class level and to remind the teachers of the goals they had set earlier.

6.2 Discussion of performances on the January-assessment

In the Netherlands, the analyses that are commonly conducted using the student monitoring systems are those analyses that predominantly focus on the development of individual students; analyses of the progress of an entire class are less frequently executed. Furthermore, adaptations to the teaching practice following the results of these analyses are still relatively uncommon (Inspectorate of Education, 2010a; Ledoux et al., 2009; Schildkamp & Kuiper, 2010). During this meeting, attention was paid to both these aspects by discussing the output of the error analysis of the January-assessment on the basis of group performance and by investigating how the analysis could be used for data-based instructional decisions. For this purpose, the homework assignments distributed during the fifth meeting were discussed in detail. As most of the participants were not familiar with the error analyses and it was only briefly discussed in the fifth meeting, its purpose and possibilities were explained in more detail. The output from a fictitious class was used to illustrate this. With respect to this fictitious example, we also provided suggestions on why and how the teacher of this class could adapt his or her whole-class teaching. When discussing their own group-error analyses, the teachers were requested to explain the over- or underachievement within specific domains. For example, if a large percentage of the class scored relatively low in the "time domain" (one of the subdomains for mathematics, in Appendix 2), teachers were asked whether this topic was sufficiently present in the curricular textbooks they used in their school and whether they had sufficiently targeted this domain during instruction. Practical tips on how to tackle such issues were provided.

6.3 Mathematical subdomains and the diagnostic conversation

As the error analysis (one of the analyses in the student assessment system) indicates students' performance on the different mathematical subdomains (see Appendix 4F), the teachers need to possess sufficient knowledge of the content and characteristics of these subdomains. We provided this information using the categorization of domains employed in the Cito assessment system (Cito, 2003). Following this information, the teachers were asked to classify several mathematical exercises into the corresponding subdomains. The correct classifications were discussed during the meeting, and a handout with additional examples per domain was distributed among the participants.

The error analysis only marks students' relative strong and weak achievements on particular mathematical domains (i.e. *what* goes right or wrong) but a more detailed analysis is needed to determine *why* students make certain mistakes. Information about omissions or misconceptions in the student's content knowledge is crucial in this respect (Van Groenenstijn, Borghouts, & Janssen, 2011). This information can be attained by means of a "diagnostic conversation". During this meeting, we stressed the importance of teacher-student interactions as these interactions help to attain valuable information on a student's knowledge and skills. When conducting a diagnostic

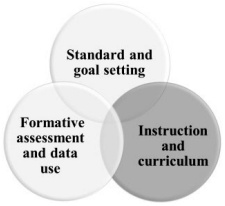
conversation, we recommended teachers to focus on the students' problem solving capabilities, because the mathematics curricula do not pay specific attention to this problem solving skill while it is an essential part of the Cito-assessments (Jacobse & Harskamp, 2011). In the problem solving process, several stages can be distinguished[33]: a) the approach phase, which refers to the steps of reading, exploring and planning, b) the calculation phase, which refers to the step in which the computational plan is implemented by the student, and c) the evaluation phase, which refers to the step in which the student evaluates the process and verifies the answer. Knowledge of students' skills in these different phases would enable the teachers to adapt their instruction in such a way that it meets their students' needs.

6.4 Resetting the performance goals

At the end of the meeting, the teachers were asked to re-examine the performance goals which they had set for the individual students in the fourth meeting. For this purpose, the teachers used the overview handed out to them earlier and discussed its content. On the basis of this empirical information the teachers could either reset (some of) their goals or leave them unaltered.

6.5 Homework assignment

In order to encourage the teachers to further investigate the output of the error analysis and to enhance their awareness of the nature of the students' mistakes - i.e., focusing on the *why* -, the teachers were asked to interview a student about the mathematical subdomain on which he or she had scored weakly. By discussing exercises with the student, the teachers were asked to assess the student's way of thinking and his/her misconceptions, and to classify these into the problem solving phase(s) of approach, calculation, or evaluation. The outcomes of this interview (also referred to as the diagnostic conversation) could be reported on a form. This form contained additional space so that teachers could write down an instructional plan on how they planned to adapt their instruction to meet this student's needs. This homework assignment was requested to be completed before the eighth meeting.



7. Reading comprehension instruction and curriculum

Summary of the seventh meeting: The goal of this meeting was to analyze the January-assessment results, and to extend the participants' knowledge of reading comprehension

[33] Problem solving is a complex activity. It requires several steps to be taken (Schoenfeld, 1992): read, analyze, explore, plan, implement and verify. In the PD program, these steps were reduced to three stages.

acquisition and the curriculum. During this session, the teachers were also asked to re-examine and, if desired, adjust their performance goals for the June-assessment.

Characteristics	
Type of meeting	Meeting at school
Time span	1.5 hours
Delivery mode	Whole-group presentation, group discussion
Hand-outs	Overview of the student test results on the January-assessment and the performance estimates for the June-assessment, overview of the reading strategies as discussed in various instructional and curricular materials, print-outs of powerpoint slides, additional reading materials, homework assignment for the school principal/internal support coordinator

7.1 Discussion of performances on the January-assessment

The performances on the reading comprehension test were addressed in a similar way as in the discussion of the mathematics results during the sixth meeting. However, the student monitoring system does not provide error analysis for reading comprehension, as in statistical analysis this skill is found to be unidimensional; an analysis of performance on (theoretical) subdomains is therefore not plausible (see Feenstra et al., 2010).

7.2 Reading comprehension; interplay among reader characteristics, text characteristics, and reading goal

In training the teachers in how to make informed instructional decisions, they were informed about the important determinants of reading comprehension achievement as comprehension stems from an active and interactive process between the *reader* (with a specific level of e.g., decoding skills, vocabulary and motivation), the specific *text* (with certain characteristics in regard to e.g., text genre, audience appropriateness and coherence), and the *goal* a reader has for that specific text (Snow, 2002; Sweet & Snow, 2003). Different reading goals require different reading strategies. We discussed these three types of characteristics and referred to observations made in the participating teachers' classrooms. For example, when discussing different genres, statement could be made such as *"I saw you addressing the distinction among different types of texts by asking your students "what is the difference between a narrative and an expository text?"*.

In addition, we discussed the two goals of formal reading comprehension instruction. These are 1) developing students' knowledge and vocabulary, and 2) teaching students how to control

their own reading processes by applying a number of reading strategies (van de Mortel & Förrer, 2010). During this meeting, the topic of reading strategies was given ample attention, as these strategies are crucial tools in helping the reader to understand a text (National Reading Panel, 2000; Pressley, 1998). The participants were provided with a handout containing five different overviews of reading strategies[34] - some more elaborate than others – and they were asked to compare these overviews and reflect on their content. To assist teachers in this task, teacher were asked which strategies they valued, which ones they used themselves and which ones they regarded as appropriate for the grades under study. The seven skills identified as appropriate for students in the second and third grades – addressed in Appendix 2 – had been translated into exemplary questions and linked to the different strategies provided in the overviews. Next, the strategies used in the teacher's own textbooks and workbooks (as analyzed by the one of the research team members) were considered and compared to those in the overviews and those mentioned in the guidelines of the Expertise Centre. The results of the homework assignment from the third meeting (determining the degree to which the students encountered the text types and question formats used in the Cito assessments) were also taken into account in this part of the meeting. Last, the importance of explicit instruction in reading comprehension strategies, i.e., explaining what they entail, why they are used and how they are used (Guthrie et al., 2004; Pressley, 1998) was addressed as it is commonly found that explicit instruction in reading comprehension is lacking (also in Appendix 2). These issues would be discussed in more detail in the next meeting. At the end of the meeting, additional materials were provided to the participants to advance their knowledge in these areas[35].

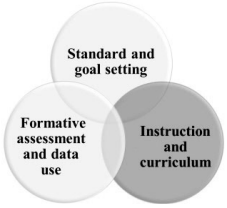
7.3 Homework assignment – school management staff

The school management staff was asked to observe a lesson in mathematics or reading comprehension given by the participating teachers, during which they had to fill in a simple observation form which targeted teachers' implementation of Direct Instruction (we specifically focused on the elements of this model pertaining to the start and end of the lesson; as discussed in Chapter 5 and aforementioned in the description of the fifth meeting). In this way, the school management staff would obtain an impression of the extent to which this model was implemented within their school's lessons. The school principal and internal support coordinator were asked to provide the teachers with constructive feedback with regard to their implementation of this practice. The observation instrument used was a simplified version of that employed by the

[34] 1) Overview of 'evidence-based' strategies by Vernooy & Stoeldraaijer, 'Algemene leerlijnen begrijpend lezen', www.taalpilots.nl 2) overview of strategies as used by the curricular textbook 'Nieuwsbegrip', 3) overview of strategies according to Filipiak (2006), 4) overview of strategies as provided by (van de Mortel & Förrer, 2010), 5) overview of strategies as used by the curricular textbook 'Kidsweek'.

[35] The additional reading material contained an analysis of the curricular text books as used in their classes (developed by Projectbureau Kwaliteit; www.schoolaanzet.nl), a special issue on reading comprehension and instruction (Loman & Marreveld, 2010), and an overview of reading strategies as provided by the curricular method Nieuwsbegrip (www.nieuwsbegrip.nl).

researchers (see Chapter 5). As the observations would be discussed during the next meeting, the assignment was requested to be completed in time.



8. Effective instruction

Summary of the eighth meeting: The goal of this meeting was to re-activate the teachers' awareness of the performance goals they had set and to train them in instructional practices which would help them to attain these goals. This was done by reflecting upon their experiences with the diagnostic conversation on mathematics, by discussing the instruction of problem solving, and by explaining and practicing how to model the application of reading comprehension strategies.

Characteristics	
Format of meeting	Meeting at school
Time span	1.5 hours
Delivery mode	Whole-group presentation, group discussion, practical modeling exercises
Hand-outs	Print-outs of powerpoint slides, overviews of goals set for mathematics and reading comprehension, modeling examples

8.1 Data use: discussion of differentiation practices

At the start of this meeting, we provided the teachers with two overviews (one for reading comprehension and one for mathematics) of the 'revised' performance goals as set in the sixth and seventh meeting. The aim of discussing this overview was threefold. First, it was used to stimulate the teachers' awareness of their performance goals. Second, the information was meant to focus teachers' attention to the performance expectations for their whole class and to draw their attention to how they could divide their class into subgroups for instruction. Third, in line with the second aim, the teachers were asked to compare the classifications in the overviews with their current ability grouping practices and to compare the differentiation practices for mathematics to those for reading comprehension. This exercise was aimed at tackling the teachers' hesitance to analyses of their students' capabilities and at fostering adaptations of their

instructional routines (recommended in, e.g., Inspectorate of Education, 2010a; Ledoux et al., 2009; Schildkamp, Visscher, & Luyten, 2009).

8.2 Instruction on problem solving and automation

As part of the homework assignment provided in the sixth meeting, the teachers were asked to conduct a diagnostic conversation with at least one of their students. During the current meeting, the information about this students' ways of thinking was discussed and classified into one of the three problem solving phases as discussed during the sixth meeting (being approach, calculation and evaluation). With respect to the first phase (the approach-phase), guidelines were given on how to teach children to approach contextual problems in a structured way. The teachers were shown practical steps which students can take when facing a word problem, such as trying to reformulate the question in their own words, underlining the relevant elements in the text, or drawing a picture representing the situation at hand (Fuchs et al., 2008; Griffin & Jittendra, 2009). With respect to the second phase (the calculation-phase) the importance of automation within the mathematics lessons was addressed. Automated knowledge of basic skills in the long-term memory facilitates more complex mathematical operations as it provides room in the working memory to conduct non-automated calculations (Ruijsenaars, Van Luit, & Van Lieshout, 2004). Especially interactive automation is a fruitful way of working, since active participation of the students increases the effectiveness of the exercises (Inspectorate of education, 2011). The researchers provided practical examples for these exercises. Last, the importance of the third phase (the evaluation-phase) was emphasized, and practical suggestions were given for how to address this phase more effectively.

8.3 Modeling strategies

Next, the concept of *modeling* was explained to the participants. Modeling is an instructional technique in which the teacher demonstrates to the students how to apply a certain reading comprehension or mathematics strategy by "thinking aloud". This approach is considered to be the primary method of showing students how they can interact with a text (e.g., Taylor & Pearson, 2002 in Fischer, Frey & Lapp, 2009) and has proven to be an effective instructional technique (Fisher et al., 2008; National Reading Panel, 2000; Pressley & Harris, 1990). This instructional practice is also discussed in Chapter 5. Teachers in Dutch primary schools are still rather unfamiliar with modeling, although this instructional approach has received some attention in journals targeting teachers and schools rather recently (e.g., Filipiak, 2006; Loman & Marreveld, 2010) and in other reading improvement PD programs (for example, Droop et al., 2012). In order to practice modeling, a hand-out was distributed in which a number of texts (from the Cito assessments of the grades under study) had been tagged with questions suitable for modeling "on the spot". We modeled three questions, after which the participants were asked to demonstrate

how they would model the other questions in their classrooms. Immediate constructive feedback was provided to the participants on their implementation of this instructional technique.

8.4 Homework assignments

To practice modeling strategies, teachers were asked to implement modeling during their book reading practices (i.e., when the teacher reads a story out loud to the class; a common practice in primary schools). The teachers were asked to prepare a section of the book that they were currently reading in their class and choose one or two reading strategies that they wanted to model. As part of the homework assignment, they were asked to adopt these strategies in the classroom and report back on their experiences. Using the same story, the teachers also had to formulate two mathematical problems which the students had to solve in pairs. In this way, the teachers could both create additional time for practicing word problems and stimulate cooperative working (Slavin & Lake, 2008). The teachers were requested to complete both parts of the assignment prior to the ninth meeting.

The second assignment concerned "near future behavior". The teachers were given a postcard containing a list of six teacher activities that were discussed during the PD program, namely: a) modeling strategies, b) interactive automation exercises, c) discussing the differences between genres, d) conducting a diagnostic conversation, e) initiating mathematical questions (following anecdotes or other situations that would lend themselves for the construction of a mathematical word problem), and f) addressing reasons for using (reading) strategies and relating the use of these strategies to topics important to the students. From these six items, the teachers were asked to mark those activities that they intended to carry out until the summer holidays. The postcard was sent to them three weeks later to remind them of these intentions.

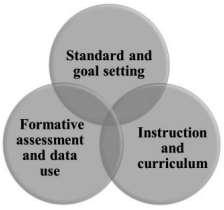
The third assignment had a similar purpose as the second one, but focused on "long term behavior". The participants were requested to write a letter to themselves about aspects of the program that they wanted to remind themselves of after the PD program had finished. We recommended the teachers to make this assignment while taking a close look at the information which was discussed throughout the school year (which they had collected in the binder). The participants were asked to hand in their letter in a closed envelope during the ninth meeting. The letters were sent to them at the beginning of the following school year to re-activate the skills and knowledge they had acquired during the PD program and to remind them of their aims.

8.5 Discussion of teacher observations – school management staff only

At this point the teachers were requested to leave the meeting. Now the school management staff was asked to report on their classroom observations, their conversations with the teachers, and the insights they had gained from their impressions and from filling in the observation instrument. These experiences were compared to our own observations of these teachers' instructional practices.

8.6 Homework assignment – school management staff only

This assignment focused on the connection between the data and the actual teaching in the classroom, building on the classroom observations that were made. The school staff was requested to analyze *the longitudinal development of the group means* for the classes observed (discussed in appendix 4H) and to compare this analysis to their perceptions during the lesson observations. As the longitudinal development of the group mean would, at least partly, give an indication of the educational quality, this assignment might be of use in helping to explain the developments in student performance. For instance, if the group mean development's learning curve was less steep than the national average, the lesson observations might help to grasp in which aspects of teacher instruction improvements should be made. In the reverse case, a very steep learning curve of the group mean might indicate "high-quality teaching" and the instructional practices of the teacher responsible for this result could then be discussed in the entire school team to stimulate other teachers to implement such high-quality teaching.



9. Evaluation of the PD program

Summary of the ninth meeting: The goal of this meeting was to connect students' performance on the June-assessment to the performance goals (in order to establish to what extent teachers had attained their goals) and to evaluate the PD program as a whole.

Characteristics	
Format of meeting	Meeting at school
Time span	1.5 hours
Delivery mode	Whole-group presentation, group discussion
Hand-outs	Print-outs of powerpoint slides, June-assessment scores, analyses of longitudinal group mean developments and cross-sections, overview of student performance in relation to performance goals

9.1 Discussion of the homework assignments

First the homework assignments of the previous meeting were briefly discussed. The teachers shared their experiences with modeling reading strategies and the formulation of mathematical

exercises. Also the execution of teachers' instructional intentions (as marked on the postcards during the previous meeting) was addressed.

9.2 Data use: student achievement in relation to the performance goals

The results of the June-assessment were discussed, as these assessments had been administered to the students prior to the current meeting. Data from the different steps of the multistep procedure as well as later meetings on student performance were combined into one overview. These overviews were handed out and discussed, while we pointed out visible patterns of over- or underperformance in relation to the teacher-set goals. The results were linked to the teachers' (assumed) level of ambition: some teachers indicated that they had been somewhat more "cautious and reserved" than other teachers in their goal setting practices. After discussing the results at the classroom level, the participants received comparisons between the performance goals and the actual achievements at the individual student level. As the focus of this meeting was particularly on evaluation at the group level, these accounts of individual students' performances were only briefly dealt with.

9.3 Data use: learning gains at the class level

Prior to the meeting, the school management staff was requested to deliver the output of the *longitudinal developments of the group mean scores* (see appendix 4H) as well as *cross-sections* (see appendix 4G) pertaining to the second- and third-grade students. The results of these data analyses were discussed with the participants. We provided support to the participants in their interpretation of the output. The development of the group mean was considered indicative of the educational quality in the different classes and facilitated the comparison of parallel classes. These analyses could be further explained by comparing them to the cross-sections, as they depict the distribution of achievement within a class. In addition, the results of the school management staff's homework assignment (in which students' performance data was connected to the teachers' observations) were discussed. The aim of combining this information with the former analyses was to foster the participants' insight into issues such as student growth, school performance, and the important role of effective teaching therein.

9.4 General remarks with respect to the lesson observations

As part of the PD program, the researchers observed one mathematics and one reading comprehension lesson of each participating teacher, both at the start of the PD program and at the end. During this evaluation meeting, general constructive feedback was given.

9.5 Evaluation of the PD program

At the end of this last meeting, the participation in the project as a whole was evaluated. The researchers asked the participants for comments and suggestions for improvement. In addition, they summarized the PD program's most important aspects of which they hoped that they would

be maintained within the school team. These were: a) a continuing focus on goals and performance, b) team discussions about data and teaching in an open atmosphere, c) teacher reflection on their own instruction practices, and d) a continuing focus on the provision of high-quality instruction in mathematics and reading comprehension. Finally, the researchers expressed their gratitude with regard to the participants' collaboration in the program.

Appendix 2: The Dutch educational context in relation to the PD program's three components

The studies presented in this dissertation were conducted to evaluate the effectiveness of a multicomponent teacher Professional Development (PD) program. This PD program aimed to improve students' reading comprehension and students' mathematics due to performance concerns for both areas (Ministry of Education, 2009; 2010). All aspects of the program which have been reported in this dissertation in relation to reading comprehension (for example, setting standards and performance goals) have also been conducted for mathematics. The content of the Appendices will pertain to the PD program as it was delivered to the participants, and will thus pertain to both subject areas. The results of the program with respect to mathematics will be discussed further in the dissertation of Ritzema (forthcoming).

In this section, several aspects of the Dutch educational context are discussed in relation to the three components of the PD program, being 1) setting standards and performance goals for every student, 2) applying formative assessment and data use, and 3) acquiring relevant instructional skills and (content and curriculum) knowledge in both reading comprehension and mathematics. We will start by addressing the recent implementation of national standards in the Netherlands, which is appropriate in light of the program's first component. After this, we discuss the Cito LOVS standardized assessment system as well as the different student monitoring systems which are used in the Netherlands. This information is relevant in light of the program's second component. In regard to the third component on relevant instructional skills and (content and curriculum) knowledge in reading comprehension and mathematics, we will give an outline of the general second- and third-grade performance expectations in these areas and address difficulties in the field of instruction which are commonly identified. General information on the curricular textbooks and alignment of these textbooks to the Cito assessments will be discussed as well.

1. Recent implementation of standards in the Netherlands

A *standard setting procedure* was conducted as part of our teacher PD program. Commonly, this procedure is conducted in countries that employ standards-based education (such as in the United States, see Cizek & Bunch, 2006) in order to identify performance categories on, e.g., state-wide or national tests (Hambleton & Pitoniak, 2006). Our reason for conducting this procedure was slightly different: the participating teachers would formulate their student-specific performance goals on the basis of the performance categories they established themselves. Recently, the political climate in the Netherlands has become more oriented toward working with standards in education. By doing so, the Netherlands have followed in the footsteps of other countries, such as the United States, England, Germany, New Zealand, Australia and South

Africa, where standards-based education has already been introduced (Fuhrman, 2001; Klieme & Maag Merki, 2008; OECD, 1995; Taylor, 2009). Commonly, two types of standards are identified. *Content standards* define what should be taught and what students should learn. *Performance standards* provide descriptions and examples of what students have to know and do, to demonstrate proficiency in the knowledge and skills framed by the content standards (Ravitch, 1995). In the Netherlands, particularly the implementation of performance standards is a relatively recent development. Content standards, known as the core objectives[36], were already actuated in 1993. In 2010, performance standards have been defined for the end of primary school in grade six and for the end of each academic track in secondary school (the Netherlands has a tracked secondary school system). By having performance standards for these points in time, it is expected that they will help boost current achievement levels as well as facilitate the transition from primary to secondary and from secondary to upper-secondary and higher education. The Ministry has identified two performance categories, namely *basic* (a performance level currently attainable for 75 percent of the student population) and *proficient* (a performance level currently attainable for 50 percent of the student population)[37]. These standards have been mandated by law for grade six since August 2010 (Ministry of Education, 2010) yet changes still need to be made in the Cito LOVS assessment system before it can be measured at which level a student is performing (e.g., 'basic' or 'proficient' when using the performance category classification). These assessment changes are scheduled for 2014-2015 (Ministry of Education, 2011). In short, the Dutch national performance standards are not yet operational and for the grades under study in this dissertation (grades two and three), no performance standards have or will be set nationally.

2. The Cito LOVS standardized assessment system and student monitoring systems used in the Netherlands

2.1. The Cito LOVS standardized assessment system

The Cito LOVS standardized assessment system, developed by the Netherlands Institute for Educational Measurement, is the most widely used assessment system in the Netherlands: it is employed in approximately 85 percent of the Dutch primary schools (Inspectorate of Education, 2010b). It provides standardized assessments throughout primary school, in different subject areas. These assessments will be referred to as the *Cito assessments*. For most subject areas, there are two yearly assessments. The midway-of-the-school year assessment is conducted in January, and the end-of-the-school year assessment is conducted in June. In the remainder of the

[36] in Dutch: kerndoelen
[37] For designing these standards a committee was installed. The standards have not yet been connected to test scores, but will be in the near future.

Appendices, these assessments will be referred to as the *January-assessment* and the *June-assessment*.

The Cito assessments for reading comprehension and mathematics cover a broad content and have a reliability of above $\alpha = 0.89$ and $\alpha = 0.91$ respectively (Feenstra et al., 2010; Janssen et al., 2010). Both assessments have been approved by the Dutch National Committee of Tests and Testing, responsible for the review of tests (COTAN). After conducting these assessments, students receive a test score which is indicative of their proficiency in either reading comprehension or mathematics. Specifically for mathematics, performance on different dimensions, or subdomains, is identified as well. These mathematical domains are slightly different across second and third grade. In second grade, three domains are distinguished. These three domains are 1) numerical knowledge, 2) arithmetics, and 3) geometry, time, and money. In third grade, four domains are distinguished being 1) numerical knowledge, 2) arithmetics, 3) geometry, and 4) time and money. For reading comprehension, no distinction in subdomains is made as psychometric analyses of performance on the reading comprehension assessments have indicated that reading comprehension is a unidimensional skill.

2.2. Student monitoring systems in the Netherlands

Test results of Cito-assessments can be registered and analyzed in administrative *student monitoring systems*. In the Netherlands, the use of student monitoring systems by teachers is still rather limited for analyzing problems and adapting instruction, and teachers who do use the student monitoring systems are often unaware of the possibilities for more sophisticated analyses (Ledoux et al., 2009; Meijer & Ledoux, 2011; Schildkamp & Kuiper, 2010; van der Kleij & Eggen, 2013). The three most commonly used student monitoring systems are the Cito student monitoring system, Parnassys, and ESIS (Meijer & Ledoux, 2011). In the Netherlands, the Cito student monitoring system is the most frequently used student monitoring tool. In our PD program, we used the Cito student monitoring system when training the participants in their use of such systems: the majority of the participants made use of this system in their school. For schools working with Parnassys and ESIS, we explained the similarities and differences between these three systems as often as possible in order to support the transfer of newly acquired knowledge and skills toward these two student monitoring systems: Parnassys and ESIS contain a majority of the analyses which are available in the Cito student monitoring systems but there are some differences in how to the results are acquired and how the output should be interpreted. In realizing this transfer of explanations and assignments from the Cito student monitoring systems to Parnassys and ESIS, we collaborated with the publishers of these three systems. One school worked with the Magister system. For this system, we could not provide instructions for transfer. However, as this school had previously worked with the Cito student monitoring system, they could easily navigate through the Magister system themselves. For the remainder of the

Appendices (for instance, in our account of the meetings on data use as reported in Appendix 2) we will exclusively refer to the Cito student monitoring system.

3. General performance expectations, common instructional practices, and the curriculum

3.1. Reading comprehension: general performance expectations

In the Netherlands, primary school starts with two years of kindergarten, followed by grades 1-6. Formal reading instruction starts in first grade. During this grade, the focus is on learning and improving the students' decoding skills. Although reading comprehension commonly starts at the beginning or halfway the second grade, there has been debate about the best time to start teaching this subject (Stoeldraijer & Vernooy, 2007). With respect to the performance expectations for second- and third-grade reading comprehension, the Expertise Centre for Dutch Language (2010) has developed guidelines on what should be expected of students during the grades under study. Throughout the PD program, these guidelines were used to help the teachers in their thinking about reading comprehension performance. The seven skills that students in the second and third grades are expected to master (as explicated in the guidelines) are the following:

1. Being able to identify the main topic of a text and to activate one's own prior knowledge on this topic.
2. Being able to connect an anaphor (a word or group of words referring back or forward to another word or group of words) to its referential tie.
3. Knowing what to do to succeed in comprehending a difficult sentence or sentences, for example, by re-reading the same section of words more slowly or looking up the meaning of a difficult word in the dictionary.
4. Being able to predict future information/content in a text.
5. Being able to process information provided in a text as well as "read between the lines".
6. Being able to distinguish among different genres, for example a narrative, expository, directive, descriptive, or argumentative text.
7. Being able to recognize the structure of a narrative text (i.e., begin, middle, and end, including the introduction of the main characters, the plot of the story, and its ending).

3.2. Reading comprehension: common instructional practices and the curriculum

In Dutch primary schools, reading comprehension lessons generally take the following sequence. First, the pupils read a text either out loud or in silence. Next, a few questions about the

text are discussed with the whole class, after which the pupils have to answer the remaining questions independently. Last, the correct answers are discussed with the whole class (Aarnoutse, 1992). During these lessons little explicit instruction is given (e.g., Van Elsäcker, 2002) and there is little differentiation between students (Van Berkel et al., 2007; Van Elsäcker, 2002). These are aspects which call for improvement. According to Collins-Block and Pressley (2002 in Houtveen, 2002), teachers do not provide instruction in reading comprehension because they are unaware that this may improve comprehension. Instead, pupils are expected to master this skill on their accord via immersion. Moreover, Stoeldraaijer & Forrer (2012) hypothesize that teachers in the Netherlands find reading comprehension a difficult subject to teach, given the complexity of the different reading comprehension skills and the curricular textbooks used, which are not always as clear as they should be. These textbooks have been criticized as being "more bulky than necessary, containing a substantial amount of material that has little or nothing to do with learning to read" (Houtveen & Van de Grift, 2012, p. 88). They also contain a large number of reading strategies, but not all of these strategies which are presented as "effective" can be supported by empirical evidence (Droop et al., 2012; Stoeldraijer & Forrer, 2012). The inadequacy of the curriculum is considered to be problematic as it is known that teachers in the Netherlands follow the content of the curricular textbooks to a very large extent in their lessons (Meelissen et al., 2012).

The content of the curricular textbooks used in second- and third-grade differ somewhat in comparison to the Cito assessment for reading comprehension in terms of text types and question types. As these textbooks differ per publisher (and there is a large number of publishers and textbooks available for teaching reading comprehension), we will only discuss characteristics of the Cito assessment here. For the grades under study, the Cito assessments contain narrative, expository, directive, and argumentative texts (Feenstra, Krom, & van Berkel, 2007a; 2007b). Especially the latter two types are relatively uncommon in the Dutch student textbooks for the grades under study. In addition, the Cito assessments make use of the following question formats: 1) multiple choice questions, 2) multiple choice items requiring the student to replace missing words in a text (i.e., a cloze test), and 3) items requiring the student to identify the first or last sentence of a short narrative text in which the sentence order has been mixed up (Feenstra, Krom, & van Berkel, 2007a; 2007b). Especially the latter two are relatively uncommon in the Dutch student workbooks.

3.3. Mathematics: general performance expectations

The formal mathematics education starts in first grade. In 2006, eleven content standards were formulated to determine the content that should be offered throughout primary school, including grade-specific elaborations (Buijs et al., 2008). The aim of these standards is to achieve

continuity in the learning progression[38]. For the grades under study (grades two and three), the following knowledge and skills were formulated (and required to be mastered)[39]:

1. Mathematical knowledge and skills
    a. Knowing the mathematical language and being able to solve practical mathematical problems (*word problems*).
2. Numerical knowledge
    a. Knowing and understanding the (structure of the) number system up to 1000 (and further).
3. Arithmetics:
    a. Being able to do additions and subtractions up to 100 (grade 4) and up to 1000 and further (fifth grade).
    b. Being able to do multiplications up to 10 and corresponding divisions (starting multiplications in fourth grade, automation in fifth grade and divisions mainly in fifth grade).
4. Geometry, time, and money:
    a. Getting acquainted with natural measures (length, weight, volume).
    b. Knowing the metric system of length (and being able to perform calculations with it).
    c. Being able to tell the time (on digital as well as on analogous clocks).
    d. Knowing the money system and being able to perform calculations with money.

3.4. Mathematics: common instructional practices and the curriculum

The Dutch curricular textbooks all contain grade-specific information on learning progression. Dutch teachers base their lessons mostly on their curricular textbooks (Meelissen & Drent, 2008). These textbooks offer advice concerning issues such as lesson structure and differentiation aspects (such as task difficulty and pacing). In Dutch primary schools, the mathematics lessons generally take the following sequence (as suggested by the textbook). First, the lesson begins with a short activating exercise. Next, the teacher provides whole-class instruction. After that, the

[38] For other references to the learning progression in the Netherlands, see van de Craats (2007) and the "rekenlijn"-website (Freudenthal Institute, SLO, & KPC, 2010).
[39] The description of the knowledge and skills required for the grades under study is rather rough and is not meant to provide a refined specification (for more information, see Buijs, Klep, Noteboom, & Klein Tank, 2008; van den Craats, 2007).

students do seatwork which entails making exercises which are based on the preceding instruction, while the teacher gives small-group instruction to the weaker performing students (also referred to as extended instruction). Then, all children do seatwork. During mathematics lessons, students thus work quite some time without instruction from the teacher (Harskamp, 2010). Several curricular textbooks dictate that in each week, two teacher- and three student-centered lessons should be provided; in the latter, the instructive role of the teacher is less dominant. As already mentioned, curricular textbooks provide suggestions for working with differences in task difficulty and for extended instruction. With respect to extended instruction, recommendations are made on additional repetition and simplification of the subject matter, using different materials to enhance the students' understanding. However, there are some doubts about the effectiveness of how these differential practices are applied in classrooms as teachers do not use the information from the student monitoring system. Not using information from this system leads to instructional practices that still do not really fit the students' needs (Inspectorate of Education, 2010a).

In the Netherlands, the learning of mathematics is nowadays more focused on contextualized items as opposed to the traditional approach which was purely based on solving arithmetic calculations. Although the modern mathematical textbooks make use of real-world questions and contexts to elicit solution strategies for improving the students' understanding, Dutch teachers' interpretation and implementation of this intended curriculum differs quite strongly compared to the way it was meant to be implemented (Royal Netherlands Academy of Arts and Sciences, 2009). Furthermore, the contextual exercises presented in these curricular textbooks seem to focus primarily on introducing a new type of problem and are designed in such a way that they facilitate the nature of the calculation which has to be made. These exercises focus mainly on improving the students' arithmetic skills and solving simple, superficial contextual problems (Royal Netherlands Academy of Arts and Sciences, 2009). Yet the question items in the Cito assessments for mathematics require both computational skills as well as more abstract 'problem solving' skills in order to understand the contextual exercise at hand. In the curricular textbooks, students are hardly expected to transfer word problems into actual sums (Meelissen & Drent, 2008). Since there is no emphasis on teaching students how to systematically solve word problems, they are bound to get easily confused when confronted with more complex word problems (Janssen et al., 2005).

Appendix 3: Overview of the cutscores and performance categories

The studies presented in this dissertation were conducted to evaluate the effectiveness of a multicomponent teacher Professional Development (PD) program in regard to reading comprehension. The teacher PD program targeted this subject area as well as the subject area of mathematics due to performance concerns for both areas (Ministry of Education, 2009; 2010). The results of the program with respect to mathematics will be discussed further in the dissertation of Ritzema (forthcoming).

Below, the cutscores and the accompanying performance categories are presented for both mathematics and reading comprehension. These cutscores were set in meeting 1 and 3 (discussed in Appendix 1 and in Chapter 3 of this dissertation).

Table 1

Overview of the cutscores for mathematics and reading comprehension

Cutscore	Mathematics Second grade	Mathematics Third grade	Reading comprehension Second grade	Reading comprehension Third grade
Minimum	37	57	-7	12
Basic	48	67	1	20
Proficient	61	77	12	30
Advanced	71	92	27	45

On the next page, an overview of the associated performance categories is provided.

Table 2

Overview of the performance categories and the associated range of assessment scores for mathematics and reading comprehension

Performance category	Mathematics Second grade	Mathematics Third grade	Reading comprehension Second grade	Reading comprehension Third grade
Below minimum	≤ 36	≤ 56	≤ -8	≤ 11
Minimum	37 – 47	57 – 66	-7 – 0	12 – 19
Basic	48 – 60	67 – 76	1 – 11	20 – 29
Proficient	61 – 70	77 – 91	12 – 26	30 – 44
Advanced	≥ 71	≥ 92	≥ 27	≥ 45

Appendix 4: Description of the data analyses used in the PD program

In this appendix, reference is made to several analyses provided by the student monitoring system. The participants received training in filling in and interpreting this system. Below, their analyses are briefly described.

During the second meeting, the participants were trained in performing the following analyses:

A. *An estimation of future performance*

Based on the prior achievements on the Cito assessments in a particular subject area, the Cito student monitoring system can provide an estimate for each student's "expected future performance". Estimates are provided with respect to two subsequent Cito assessments.

B. *A performance overview (in the PD program: from the previous school year)*

The Cito student monitoring system can list the performances of students on the January and June-assessments for each school year.

C. *A progress report* (*'ability growth report'*)

The Cito student monitoring system can calculate the difference between a current and a prior assessment score to see whether or not a student has improved his or her performance over a certain time period.

During the fifth meeting, the teachers and management staff received training in performing analyses relevant to their function. The *teachers* were trained in the following analyses:

D. *The alternative student report (a performance comparison which makes extreme high or low scores more tangible)*

This analysis gives information about the actual performance level of students with extreme high or low scores. As the proficiency scale for the Cito-assessments is the same for grades one to six, this analysis can indicate the actual level in terms of a grade mean that is (far) below or above the student's current grade. For example, a second-grade student with a very high math score might perform similar to the average performance of Dutch students in fourth grade (which is two years above the mean).

E. The average performance of the class

The Cito student monitoring system can produce a graph representing a particular class' average test performance for several consecutive school years. Salient patterns signal a need for further investigation.

F. Error analysis for the mathematics assessment ('*category analysis*')

The Cito-assessments for mathematics can be further analyzed by calculating students' weighed performance on the different mathematical domains. The analysis compares a student's actual performance on a specific mathematical domain to his/her expected performance, relative to his/her overall performance level. This comparison is made for all domains assessed, resulting in an overview of how the student scores on each mathematical domain relative to his overall performance level. The analysis delivers two tables. The first one shows how the actual performance deviates from the expected performance per student. This leads to three categories, for which different signs are given 'non-salient', 'salient', and 'very salient'. These categories provide information on whether and how much the actual performances deviate significantly from the expected achievements. In the second table, group information is summarized as regards positive and/or negative deviations of the group's expected performance on the different mathematical domains. On both levels - individual and whole group - the teachers are thus informed about students' strengths and weaknesses in domain-specific knowledge and skills.

During the fifth meeting, the *school management staff* was trained in performing the following analyses:

G. Cross-section of performance levels

Students' performance on the Cito assessments can be classified in terms of ability level indicators (A to E, and I to V): we refer to the A to E distribution as it is still the most common classification in Dutch primary schools. The top 25 percent of the performance distribution (thus, the best performing students) are given an 'A'. The next 25 percent of the performance distribution (thus, the 'second best' performing students) receive a 'B'. The next 25 percent of the performance distribution (thus, the 'third-best' performing students) receive a 'C'. The 15 percent of students performing below that C receive a 'D' (15%), and the 10 percent of the lowest performing students receive an 'E'. This distribution is computed by the Cito student monitoring system. The cross-section analysis has two options for reporting its results: 1) cross-sections for one or more classes with respect to one subject area, and 2) cross-sections for one or more classes with respect to different subject areas.

H. Longitudinal developments of the group's mean score(s) of one or more classes ('trend analysis').

The Cito student monitoring system provides two options for looking at the development of student performance throughout primary school with the help of means: 1) the cross-sectional development of the group mean, and 2) the longitudinal development of the group mean. The cross-sectional development of the group mean shows the scores of several grades (e.g., grades one, two, and three) on a specific subject area throughout different years. Here the performance stability of different groups of students is thus compared. The longitudinal development of the group mean indicates the stability of a specific group of students' mean growth during primary school (grade 1-6). Here the mean growth of one or more classes in one year is compared to its/their growth in other school years.

I. Mean score of a specific grade throughout several years

The student monitoring system can also show the stability of the mean scores of a particular grade throughout several years.

J. Group analyses

In this analysis, data from the cross section and the trend analysis are combined, revealing the development pattern of a class with respect to different subject areas over several years. It provides an overview of a class's mean score development, based on the five ability levels as identified in the Cito LOVS assessment system. This analysis helps to gain an insight into the developmental patterns of a class in different subject areas, again showing deviations if there are any.

Appendix 5: Lessons learned from the pilot

The PD program was conducted in the school year of 2011-2012, after the PD meetings and the associated materials had been piloted in the previous school year of 2010-2011. The experiences of the pilot study led to three relatively large modifications of the PD program's set-up. The first one pertained to the simultaneous focus on the subject areas mathematics and reading comprehension. During the pilot study, there was a five-month period for reaching the goals for reading comprehension, followed by a comparable five-month period for mathematics. However, we experienced that this five month time frame per subject area was "too short" for the teachers to be able to work toward the full attainment of their goals. Therefore, it was decided to focus on both subject domains simultaneously in the 2011-2012 PD program. After the standard setting procedures for both subjects were completed, the teachers set their goals in November/December 2011 for the June 2012 assessment, which gave them more time to work toward their goal attainment.

The second modification in the 2011-2012 PD program's set-up applies to the role of the school management staff. During the pilot study, teachers, school principals and internal support coordinators participated in the meetings, whereas the focus was mostly on the teachers' knowledge, practices and assignments. The school management staff from the pilot study schools indicated that they would have liked to receive more information and assignments targeted at their specific function in the school. Furthermore, it was found that in those schools in the pilot study where the school management staff was particularly dedicated to the project, the teachers were too. This is in line with the findings of, for example, research on data use (Schildkamp & Kuiper, 2010; Sutherland, 2004). In the 2011-2012 PD program, the content of the performance data analyses was therefore adapted in such a way that the school management staff was targeted more actively (resulting in different data analysis assignments and the use of parallel sessions during the fifth meeting). Furthermore, in the 2011-2012 PD program, the school management staff was asked to conduct a lesson observation and provide teachers with constructive feedback on the implementation of instructional practices.

The third modification concerned the instructional practices addressed during the meetings to help the teachers meet their achievement goals. In the pilot study, both modeling and the Direct Instruction model were discussed around February. However, the researchers experienced that the participants failed to recall these instructional practices later on and that their actual implementation was insufficient. Throughout the 2011-2012 PD program, both the content and value of the Direct Instruction approach and explicit strategy-instruction (modeling) were therefore more explicitly and repeatedly discussed. In addition, the teachers were given feedback

on both their modeling practices (provided by the researchers) and on their implementation of Direct Instruction (provided by the researchers as well as the school management staff).

All modifications to the PD program's set-up following the pilot study aimed to better fit the needs of the participants in order to increase student achievement.

Appendix 6: Evaluation form used during the standard setting meeting

Evaluation form

Date: ……..

Name: …….

1. Did you find today's meeting …

…clear	0 yes	0 a little	0 no	Comment: ……..
…useful	0 yes	0 a little	0 no	……
…informative	0 yes	0 a little	0 no	..……
				……

2a. Was the explanation of the standard setting procedure …

…clear	0 yes	0 a little	0 no	Comment: ……...…
…useful	0 yes	0 a little	0 no	……
…informative	0 yes	0 a little	0 no	……
				……

2b. Was it … to you how to carry out round one (the individual round)?

…clear	0 yes	0 a little	0 no	Comment: ……...…
…useful	0 yes	0 a little	0 no	……
…informative	0 yes	0 a little	0 no	……
				……

2c. Was it … to you how to carry out round two (the discussion round)?

…clear	0 yes	0 a little	0 no	Comment: ……...……
…useful	0 yes	0 a little	0 no	……
…informative	0 yes	0 a little	0 no	……
				……

2d. Was it … to you how to carry out round three (the empirical data round)?

…clear	0 yes	0 a little	0 no
…useful	0 yes	0 a little	0 no
…informative	0 yes	0 a little	0 no

Comment: ……..……
……
……
……

2e. Do you find your own cutscores from … to be well-considered?

…round 1	0 yes	0 a little	0 no
…round 2	0 yes	0 a little	0 no
…round 3	0 yes	0 a little	0 no

Comment: ……...…
……
……
……

3. Do you think that your participation in today's meeting will influence your teaching and/or will lead to an improvement of your teaching behavior?

0 yes	0 a little	0 no

Comment: ……
……..
……..

Remarks and/or suggestions:

Tip: *what could have been improved in today's meeting?* … …….
……..
……..

Top: *what did you like or find interesting in today's meeting?* ……
……..
……..

Thank you for completing this form!

Appendix 7: Observation instrument

The time-sampling instrument consisted of several aspects of teachers' and students' activities which were coded every two minutes for the entire lesson (maximum duration: 60 minutes). With help of this instrument, we coded the phase of the lesson as well as several aspects of teacher behavior, namely teacher position, teacher talk, and the student at whom this teacher talk is directed. These categories are elaborated in Table 1. Furthermore, we selected four specific students prior to the lesson observation on the basis of their academic performance and coded the activity in which they were participating. These four students, being the same at pre- and postmeasurement, were a) a very weak performing student, b) a weak performing student, c) an average performing student, and d) an advanced student. The teachers were not informed that we were interested in students' activities nor were they informed which students were selected for the purpose of the observation.

In Table 2, we present the scheme we used for conducting the time-sampling data collection. In light of parsimony, only the first fifteen two-minute intervals (i.e., the first 30 minutes) and the last interval (at the 60^{th} minute) are depicted in this table.

In Table 3, the high-inference measure is presented. This measure contained 16 items pertaining to different aspects of teachers' behavior which were filled in directly after the lesson was observed. All items had dichotomous response options, with a 0 (*no*) or a 1 (*yes*) in the case that implementation of these practices was observed.

Table 1

Time-sampling categories

Variable	Categories	Explanation
Classroom organization (i.e., Phase)	1. Whole-class instruction	• The teacher provides whole-class instruction.
	2. Extended instruction	• One or more students receive extended instruction, while the other students in class do seatwork.
	3. Seatwork	• All students are working on exercises (individually, in pairs or small groups).
Position of the teacher	1. In front of the class	• The teacher is standing or sitting in front of the students
	2. At a student's table or group of tables	• The teacher is standing or sitting at a student's table or a group of tables
	3. Walking around	• The teacher is walking around the class
	4. At the desk	• The teacher sits at the desk
	5. Other	• Teacher position is not options 1 - 4 (e.g., outside the class)
Teacher talk	1. Task at hand	• The teacher refers to the task at hand *(e.g. 'we will start with exercise 1, page 14', 'now, we are going to do some automation exercises')*
	2A. Explanation pertaining to content	• The teacher provides information on the task, strategies and solutions *(e.g. 18 times 6. To solve this, you can take two steps. First, you calculate 10*6 and then 8*6')*
	2B. Content-related questioning	• The teacher asks for information on the task, strategies and solutions *(e.g. 'How much is 6*8?' or 'How did you answer that question?')*
	3. Organization	• The teacher refers to the general sequence of the lesson or to conditions for working *(e.g. 'Maria, please pay attention' or 'you can come to me after the whole class instruction')*
	4. Other	• Teacher talk is not options 1 - 3 (e.g., '*well done*').
Student who is addressed during teacher talk	1. Very weak achieving student	• Selected student, minimum level (10% lowest performers)
	2. Weak achieving student	• Selected student, basic level (25% lowest performers)
	3. Average achieving student	• Selected student, proficient level (average performer)
	4. Advanced student	• Selected student, advanced level (25% highest performers)
	5. Other student	• Non-selected student in the same class
	6. Group of students or whole class	• The whole class or a (small) group
	7. Other	• Other, e.g., a colleague who comes into the class or students of the other grade in a multi-grade class
Activity very weak student (the same for the weak, average, and advanced student)	1. Whole-class teaching	• The student is engaged in whole-class teaching
	2. Extended instruction	• The student receives additional instruction in a small group
	3. Individual teacher instruction	• The student receives additional, individual instruction or is working individually with the teacher
	4. Working independently	• The student works on his own or with (a) peer(s)
	5. Other	• The student is outside the classroom or is working on exercises from a different subject area.

Table 2

Time-sampling scheme

	Class	Teacher			Student				
Time	**Phase**	**Position**	**Teacher talk (TT)**	**Student addressed during TT**	**Very weak**	**Weak**	**Average**	**Advanced**	**Elaboration**
0									
2									
4									
6									
8									
10									
12									
14									
16									
18									
20									
22									
24									
26									
28									
30									
…									
60									

Table 3

High-inference measure

	The teacher ….	**No**	**Yes**
1	… summarizes the content of the prior lesson or activates relevant prior knowledge		
2	… explicates the learning goal, content and/or topic of that lesson		
3	… starts the mathematics lesson with an automation exercise		
4	… clarifies in which way the assignments are accomplished in a satisfactory manner		
5	… provides extended instruction to a student or group of students		
6	… differentiates for the weaker students in the assignments these students are expected to complete		
7	… differentiates for well achieving students in the assignments that these students are expected to complete		
8	… lets students that have completed their assignments … - for math: work on more advanced mathematical materials - for reading: read for themselves		
9	… gives the students time to answer the question (+ 3 seconds).		
10	… repeats the right answer.		
11	… compliments students when they have answered the question correctly.		
12	… At the end of the lesson, the teacher returns to the learning goal of that lesson and/or the new skill and/or knowledge that have been addressed		
13	… connects the content of the current lesson to the following lesson		
14	… goes into depth in regard to the approach and used strategies after right answers are provided		
15	… goes into depth in regard to the approach and used strategies after wrong answers are provided		
16	.. models his/her application of knowledge, skill or strategy by thinking aloud		

Samenvatting (Dutch summary)

1. Introductie

Basisscholen hebben de belangrijke taak om leerlingen toe te rusten met voldoende leesvaardigheid ten behoeve van hun verdere schoolloopbaan en participatie in de maatschappij (Kirsch, 2002; Reis, McCoach, Little, Muller, & Kaniskan, 2011; Snow, Burns, & Griffin, 1998; Van Elsäcker, 2002). Momenteel zijn er echter zorgen over de leesvaardigheid van Nederlandse leerlingen; dit naar aanleiding van tegenvallende prestaties op zowel nationale als internationale leestoetsen (Ministerie van Onderwijs, 2008; 2010). Deze tegenvallende prestaties worden toegeschreven aan verschillende oorzaken, onder andere a) het gebrek aan duidelijke prestatiedoelen waar scholen en leerkrachten zich op kunnen richten (Inspectie van het Onderwijs, 2011; Ministerie van Onderwijs, 2010; Onderwijsraad, 2007), b) de beperkte mate waarin leerkrachten differentiëren tijdens de lessen begrijpend lezen (Inspectie van het Onderwijs, 2012; Van Berkel et al., 2007; Van Elsäcker, 2002), c) de beperkte mate waarin leerkrachten expliciet instructie geven in begrijpend lezen (Aarnoutse & Weterings, 1995; de Jager et al., 2002; Van Elsäcker, 2002) en d) de tekortkomingen in de begrijpend leesmethoden die in Nederland worden gebruikt (Droop et al., 2012; Houtveen & Van de Grift, 2012; Stoeldraijer & Forrer, 2012).

Met het oog op het verbeteren van de begrijpend leesresultaten hebben wij een nascholingsprogramma voor basisschoolleerkrachten ontwikkeld. De verwachting was dat leesprestaties zouden verbeteren wanneer instructie beter zou aansluiten bij de behoeften en capaciteiten van leerlingen en wanneer de instructie doelgerichter en meer expliciet zou zijn. Om deze gewenste verbetering in instructie te kunnen bewerkstelligen is er in het nascholingsprogramma gebruik gemaakt van drie componenten, genaamd 1) leerstandaarden en prestatiedoelen voor iedere leerling, 2) opbrengstgericht werken en 3) vakspecifieke kennis en instructiepraktijken. Deze drie componenten zijn geïntegreerd in één (synergetisch) pakket met de naam ‘Streef Middenbouw’ gericht op leerkrachten van groep 4 en 5. Deze doelgroep was geselecteerd vanwege het belang van goede leesprestaties op jonge leeftijd (zie, o.a., Bodovski & Youn, 2011; Snow et al., 1998) en het feit dat uit de literatuur over onderwijseffectiviteit is gebleken dat ‘wat werkt’ vaak de meeste leerwinst oplevert wanneer dit wordt toegepast bij jongere leerlingen (zie bijvoorbeeld Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988).

In de huidige dissertatie is verslag gedaan van de effectiviteit van het nascholingsprogramma. De volgende onderzoeksvraag stond hierbij centraal: halen leerlingen hogere prestaties voor begrijpend lezen nadat hun leerkrachten aan het nascholingsprogramma ‘Streef Middenbouw’ hebben deelgenomen, en kan er empirisch bewijs gevonden worden voor de verschillende onderliggende assumpties van het programma?

1.1. Opzet van het nascholingsprogramma

Van elk van de drie componenten uit het nascholingsprogramma was bekend dat deze positief samenhing met prestaties van leerlingen (voor het werken met doelen, zie o.a. Fuchs, Fuchs, & Deno, 1985; voor opbrengstgericht werken, zie bijvoorbeeld Carlson et al., 2011; voor expliciete instructie in begrijpend lezen, zie o.a. Andreassen & Braten, 2011). De drie componenten van het programma waren als volgt opgezet: in het kader van de eerste component hebben leerkrachten leerstandaarden geformuleerd voor leerlingen van verschillende niveaus. Vervolgens hebben leerkrachten op basis van deze leerstandaarden doelen gesteld voor hun eigen leerlingen, aangezien het werken met doelen de aandacht richt op het behalen van de gewenste uitkomsten (Locke & Latham, 1990). In het kader van de tweede component zijn de leerkrachten getraind in het werken met en interpreteren van toetsgegevens van leerlingen uit het leerlingvolgsysteem, iets waar nog veel winst te behalen valt (zie bijvoorbeeld Van der Kleij & Eggen, 2013). Hierbij richtte de nascholing zich op het beter inzicht krijgen in het niveau van leerlingen, zodat het onderwijs hierop aangepast kon worden (Black & Wiliam, 1998). Het was de verwachting dat differentiatie gestimuleerd zou worden door de combinatie van het stellen van doelen en het monitoren van de voortgang van leerlingen; tijdens de nascholing is regelmatig aandacht aan dit onderwerp besteed. In het kader van de derde component zijn leerkrachten getraind in het implementeren van het Directe Instructie model, een effectief leerkracht gestuurd lesmodel (Borman, Hewes, Overman, & Brown, 2003) waarbij expliciet aandacht is voor het doel van de les. Ook zijn leerkrachten getraind in het modelleren; dit is een effectieve instructiemethodiek waarbij de leerkracht hard-op denkend voordoet hoe een leesstrategie kan worden toegepast of hoe een vraagstuk kan worden opgelost (zie Fisher et al., 2008). Tot slot is er besproken wat er van leerlingen in de middenbouw verwacht mag worden op het gebied van kennis en vaardigheden voor begrijpend lezen. Dit is onder andere gedaan door de begrijpend leesmethoden die op de scholen werden gebruikt te vergelijken met andere bronnen en richtlijnen (waaronder die van het Expertisecentrum Nederlands, 2010).

Het Streef Middenbouw-project was een van de deelprojecten in het grootschalige ‘Streef’-project dat in Noord-Nederland is uitgevoerd; onder deze noemer zijn meerdere schoolverbeterings- en onderzoeksprojecten (met thema’s waaronder opbrengstgericht werken en het bepalen van leerstandaarden) in verschillende jaargroepen uitgevoerd. Toetsgegevens van leerlingen zijn door de hele school verzameld: hierdoor konden de prestaties van leerlingen in ‘onbehandelde’ klassen als vergelijkingsmateriaal dienen voor de ‘behandelde’ klassen.

In het schooljaar van 2010-2011 was er sprake van een pilotjaar: in dit jaar hebben zes scholen deelgenomen aan een proefversie van het nascholingsproject. In het schooljaar 2011-2012 hebben 19 scholen uit Noord-Nederland deelgenomen aan de definitieve versie van het Streef Middenbouw-project. In dat schooljaar hebben 33 leerkrachten het nascholingsprogramma gevolgd; deze leerkrachten gaven les aan 451 leerlingen. Ook de Intern Begeleiders (IB’ers) en

schooldirecteuren van de betreffende scholen hebben meegedaan aan het nascholingsprogramma. Het programma heeft ongeveer 40 uur tijd in beslag genomen - verspreid over het hele schooljaar - waarin van de deelnemers werd verwacht dat zij aanwezig waren bij de negen naschoolse bijeenkomsten (duur per bijeenkomst: 1,5 tot 2,5 uur) en de bijbehorende huiswerkopdrachten maakten. Tijdens de negen bijeenkomsten, waarvan vier bovenschools plaatsvonden en vijf op de eigen school, zijn presentaties van de onderzoekers afgewisseld met interactieve werkvormen en opdrachten. Daarnaast zijn de leerkrachten driemaal geobserveerd: tweemaal door de onderzoekers (een voor- en nameting), en eenmaal tussentijds door de directeur of IB'er. De informatie uit de observaties is gebruikt om de inhoud van de bijeenkomsten meer te laten aansluiten bij het handelen van de leerkrachten, en om leerkrachten constructieve feedback te geven over hun eigen instructiepraktijk. De opzet van het programma kwam - op deze en op andere punten - tegemoet aan de aanbevelingen uit de literatuur over effectieve nascholing van leerkrachten (zie bijvoorbeeld, Garet, Porter, Desimone, Birman, & Yoon, 2001; Yoon, Duncan, Lee, Scarloss, & Skapley, 2007).

2. Samenvatting van de onderzoeksresultaten

In hoofdstuk 2 zijn de effecten van het programma op de prestaties van leerlingen onderzocht. In hoofdstukken 3 en 4 is er in meer detail gekeken naar de prestatiedoelen die de leerkrachten voor hun leerlingen hebben gesteld. In hoofdstuk 5 is de implementatie bestudeerd van verschillende instructiepraktijken die tijdens het programma werden behandeld. Een samenvatting van de resultaten zal hieronder worden gegeven.

2.1. Het effect van de leerkrachtnascholing op de begrijpend leesprestaties van leerlingen

In hoofdstuk 2 is het effect van het programma op leesprestaties onderzocht. Uit een grotere poel van mogelijke controle klassen (alle 'onbehandelde' klassen in het grootschalige Streef-project) zijn met behulp van de *propensity score matching* techniek (Rosenbaum & Rubin, 1985) de meest vergelijkbare klassen geselecteerd om te dienen als vergelijkingsmateriaal voor de klassen in de experimentele conditie. Leerlingen in de experimentele conditie bleken significant beter te presteren dan leerlingen in de gematchte controleconditie op de Cito-toets voor begrijpend lezen met een effectgrootte van $d = .37$ (en een bijbehorend 90% betrouwbaarheids-interval van $d = .20$ tot $d = .55$). De robuustheid van de resultaten is nagegaan door middel van drie variaties op het gebruikte statistische model. Ook met deze alternatieve berekeningen bleek het programma effectief, zij het met iets kleinere effectgroottes (respectievelijk $d = .29$, $d = .30$ en $d = .31$). Volgens de richtlijnen van Cohen (1988) kunnen deze effectgroottes geïnterpreteerd worden als kleine tot middelgrote effecten. Verder is onderzocht of het effect van het programma op prestaties afhankelijk was van de jaargroep waar de leerlingen in zaten en hun eerdere prestaties voor begrijpend lezen. Deze differentiële effecten bleken niet significant. Met andere woorden, alle leerlingen - onafhankelijk of ze in groep 4 of in groep 5 zaten en of ze op een

eerder meetmoment juist laag of hoog presteerden op begrijpend lezen - schenen evenveel geprofiteerd te hebben van de deelname van hun leerkracht aan het nascholingsprogramma.

2.2. Werken met prestatiedoelen

Vervolgens is er gekeken naar de doelen waarmee leerkrachten hebben gewerkt. De doelen hebben namelijk een belangrijke rol gespeeld tijdens het Streef Middenbouw-project. Een specifieke eigenschap van deze doelen was dat ze waren geformuleerd in termen van prestatiecategorieën; deze worden ook wel leerstandaarden genoemd. De categorieën (onder minimum, minimum, fundamenteel, streef en gevorderd) waren vastgesteld op de schaal van de Cito-toets voor begrijpend lezen. Door onderscheid te maken in verschillende prestatiecategorieën konden doelen worden gezet die tegemoet kwamen aan verschillen tussen leerlingen ('Aan het eind van het schooljaar, op de Cito-toets voor begrijpend lezen, wil ik dat Milan presteert op het *fundamentele* niveau en Anne op het *gevorderde* niveau'). Gedurende het schooljaar is veelvuldig naar de prestatiedoelen gekeken en verwezen aangezien wij door middel van de tweede en derde component van het programma (over opbrengstgericht werken en vakspecifieke instructie) het behalen van deze doelen wilden faciliteren. Aan het eind van het schooljaar is nagegaan in hoeverre de leerkrachten de doelen voor hun eigen klas hadden behaald.

In hoofdstuk 3 is dieper ingegaan op de leerstandaarden waarop de doelen waren gebaseerd. De deelnemers werden gevraagd om, aan de hand van een zogenoemde standaardbepalingsprocedure met hierin meerdere rondes, verschillende Cito-toets opgaven te bestuderen en aan te geven in hoeverre deze opgaven beheerst moesten worden door leerlingen van verschillende niveaus. Het doel van de procedure was het onderscheiden van de vijf prestatiecategorieën, en deze vijf categorieën zijn zowel voor groep 4 als voor groep 5 bepaald. In de standaardbepalingsprocedure ging het specifiek om het identificeren van de grens tussen twee opeenvolgende niveaus. Doordat de opgaven gekoppeld waren aan scores op de vaardigheidsschaal van de Cito-toets van begrijpend lezen, konden vervolgens elk van de vijf prestatiecategorieën gekoppeld worden aan een interval op deze vaardigheidsschaal (door afname van de Cito-toets aan het einde van het schooljaar kon worden nagegaan of de leerlingen de doelen hadden bereikt, door de behaalde vaardigheidsscores op de toets te vergelijken met de vooraf geselecteerde intervallen). Een aanname in het programma was dat de grenzen tussen de verschillende leerstandaarden - in het Engels *cutscores* genoemd - accuraat waren, en naar deze aanname is empirisch onderzoek uitgevoerd. Om de accuraatheid van *cutscores* te beoordelen is het gebruikelijk om te kijken naar het bewijs voor verschillende typen validiteit (Cizek & Bunch, 2006; Hambleton & Pitoniak, 2006; Kane et al., 1999; Norcini & Shea, 1997; Pant et al., 2009), en in dit hoofdstuk is daarom gekeken naar de procedurele validiteit en de interne validiteit van de *cutscores*. De procedurele validiteit is onderzocht met behulp van de feedback van de deelnemers op het gebied van a) duidelijkheid van de procedure, b) uitvoerbaarheid van de procedure en c) de mate waarin de deelnemers hun eigen oordelen weloverwogen vonden. De

interne validiteit is bestudeerd door de variatie in *cutscores* over de verschillende rondes in de procedure te onderzoeken: hierbij is er specifiek gekeken naar d) de aanpassingen over de rondes heen, e) de overeenstemming tussen de cutscores en empirische informatie, en f) de overeenstemming tussen de deelnemers. De onderzoeksgegevens bleken voldoende ondersteuning te bieden voor beide typen validiteit.

In hoofdstuk 4 zijn de doelen die de leerkrachten hebben gezet voor hun leerlingen nader onderzocht. Om leerkrachten te ondersteunen in het stellen van doelen voor elk van hun leerlingen hebben wij een meertrapsprocedure ontwikkeld. In deze procedure werden de leerkrachten gevraagd om eerst een inschatting te maken van een geschikt doel voor elke leerling. Vervolgens was er expliciet aandacht voor de prestaties van deze leerlingen en vond er overleg met collega's plaats. Aan het eind van de meertrapsprocedure werden de definitieve doelen voor de leerlingen gezet. In het hoofdstuk is het gebruik van deze procedure geëvalueerd door te kijken naar de aanpassingen over de verschillende rondes in de procedure heen; een gebrek aan aanpassing werd namelijk gezien als onvolkomen gebruik van de procedure. In de analyses is een significante mate van aanpassing gevonden. Vervolgens is de relatie tussen de doelen en de prestaties van de leerlingen onderzocht. Om dit te onderzoeken is er gekeken naar de mate waarin de leerlingen de doelen hebben behaald. Tevens is onderzocht of de doelen een significante voorspeller van prestatie waren terwijl er voor belangrijke leerling- en klaskenmerken werd gecontroleerd. Aan het eind van het schooljaar bleek dat bijna tachtig procent van de leerlingen op of hoger dan het gewenste niveau presteerde. Daarnaast bleken de doelen significante voorspellers van het niveau van begrijpend lezen. Hoge doelen hingen samen met hogere prestaties, een resultaat dat aansloot bij de bevindingen uit de literatuur over het werken met doelen (Locke & Latham, 1990; 2002) en leerkrachtverwachtingen (Jussim & Harber, 2005; Rosenthal & Jacobson, 1968). Tot slot bleek dat leerlingen die aan het eind van het voorgaande schooljaar lager presteerden op de Cito-toets voor begrijpend lezen extra profiteerden van een relatief hoger doel, tevens een bevinding die aansloot bij de literatuur (Good & Brophy, 2003).

2.3. Implementatie van gestimuleerde instructiepraktijken

In hoofdstuk 5 is gekeken naar de instructie die door de deelnemende leerkrachten is verzorgd. Hierbij is specifiek gekeken naar de implementatie van Directe Instructie, modelleren en differentiatie, gezien het feit dat leerkrachten in deze instructiepraktijken zijn getraind. In een vergelijking tussen de voor- en nameting kwam naar voren dat er significant meer leerkrachten modelleerden aan het eind van het schooljaar dan aan het begin van het schooljaar. De leerlingen wiens leerkrachten modelleerden presteerden significant beter dan leerlingen wiens leerkrachten niet modelleerden; de grootte van dit effect was $d = .24$ (met een bijbehorend 90% betrouwbaarheidsinterval van $d = .03$ tot $d = .46$). Wel moet worden gezegd dat het totaal aantal leerkrachten dat modelleerde op de nameting vrij beperkt was (in totaal modelleerden negen

leerkrachten op de nameting). De implementatie van zowel het Directe Instructie model als differentiatie bleek gering te zijn en veranderde niet van voor- naar nameting.

3. Discussie

Bij het interpreteren van de resultaten die in deze dissertatie zijn gepresenteerd moeten enkele beperkingen in acht worden genomen: hier zullen de drie belangrijkste worden besproken. Ten eerste is geconcludeerd dat het programma effectief was in het verbeteren van de begrijpend leesprestaties van leerlingen, maar de vraag *hoe* deze prestatieverbetering is bewerkstelligd is deels onbeantwoord gebleven. De implementatie van Directe Instructie, modelleren en differentiatie leek beperkt, zoals gemeten met het gebruikte observatie-instrument. De toepassing van deze instructiepraktijken kan dus niet gebruikt worden als verklaring voor het positieve effect van het programma op leerlingprestaties. Het gebruik van een uitgebreider observatie-instrument, met hierin aandacht voor meer algemene aspecten van instructiekwaliteit en ruimte om kwaliteitsverschillen beter van elkaar te kunnen onderscheiden, was in dit geval wenselijk geweest.

Omdat deelname aan het Streef Middenbouw-project plaatsvond via zelfselectie en de deelnemende leerkrachten wisten dat de effecten van het programma geëvalueerd zouden worden zou het *Hawthorne effect* (een fenomeen waarbij deelnemers hun gedrag verbeteren simpelweg omdat ze weten dat ze bestudeerd worden, en niet vanwege de inhoud van het programma, Shadish et al., 2002) een mogelijke alternatieve verklaring zijn voor het positieve effect van het programma op leerlingprestaties. Echter, in hoofdstuk 2 is de waarschijnlijkheid van deze alternatieve verklaring al in twijfel getrokken. Ondanks het feit dat leerkracht- en schoolkenmerken niet konden worden meegenomen in de constructie van de *propensity score* (vanwege non-response op een vragenlijst) verwachten wij dat redelijk vergelijkbare scholen en leerkrachten hebben deelgenomen aan zowel de experimentele conditie als de controleconditie; dit omdat de scholen in de controleconditie met andere jaargroepen meededen aan schoolverbeteringsprojecten gericht op redelijk vergelijkbare thema's. Op alle scholen binnen het grootschalige Streef-project zijn door de hele school toetsgegevens verzameld ten behoeve van de evaluatie van de verschillende deelprojecten, en alle leerkrachten zijn hierop geattendeerd. Desalniettemin zou een replicatie van het onderzoek met aselecte toewijzing aan condities een waardevolle aanvulling zijn op de hier gerapporteerde resultaten. Op die manier kan met meer zekerheid worden gezegd dat het Streef Middenbouw-programma het positieve effect op de leesprestaties heeft veroorzaakt.

Het derde punt dat in het kader van de beperkingen wordt besproken heeft betrekking op het feit dat de ontwikkelaars van het programma zelf de evaluatie hebben uitgevoerd. Het voordeel van deze opzet was dat de verzamelde observatiegegevens gebruikt konden worden om de inhoud van de bijenkomsten beter aan te laten sluiten bij het niveau van de deelnemers. In het licht van

mogelijke *experimenter bias* (zie Rosenthal & Fode, 1963) wordt echter aanbevolen om bij vervolgonderzoek externe onderzoekers in te schakelen; zo kan de objectiviteit van de resultaten beter gegarandeerd worden.

4. Aanbevelingen voor verder onderzoek en praktische implicaties

Voor verder onderzoek wordt aanbevolen om de aandacht te richten op het effectiever en efficiënter maken van het nascholingsprogramma. Dit kan worden gedaan door voorafgaand aan de effectmeting meer in te zetten op de implementatie van Directe Instructie, modelleren en differentiatie. Het effect van het programma op leerlingprestaties zal naar verwachting sterker zijn wanneer deze (effectieve) praktijken beter geïmplementeerd worden door de deelnemende leerkrachten. Enkele aanpassingen in het programma zullen noodzakelijk zijn om deze gewenste implementatie te kunnen bewerkstelligen, bijvoorbeeld door meer tijd uit te trekken voor de gewenste gedragsverandering, door het aantal observatie-momenten te vergroten zodat leerkrachten vaker feedback ontvangen en door handreikingen voor aanpassingen in de leesmethode aan te leveren. Vanwege het vermoeden dat de attitude van leerkrachten een rol heeft gespeeld in de beperkte implementatie van onder andere het Directe Instructie model wordt aanbevolen om, in de aanvangsfase van het vervolgonderzoek (gericht op een sterkere implementatie van Directe Instructie, modelleren en differentiatie), middels een exploratieve aanpak te kijken naar de aspecten die de implementatie van deze instructiepraktijken mogelijk belemmeren. Wanneer de oorzaken voor beperkte implementatie beter geïdentificeerd worden, kunnen deze later in het programma ook beter aangepakt worden. Tegelijkertijd is het wenselijk om naast de aandacht voor de implementatie van Directe Instructie, modelleren en differentiatie ook te kijken naar andere veranderingen in instructie als gevolg van het programma. Tijdens het programma zijn namelijk onderwerpen behandeld zoals *evidence-based* leesstrategieën en belangrijke kernconcepten voor begrijpend lezen in de middenbouw, als ook het opbrengstgericht werken binnen de les. Wellicht hebben deze onderwerpen de kwaliteit van instructie op een dusdanige manier beïnvloed dat daardoor het gevonden effect verklaard kan worden (bijvoorbeeld doordat de les rijker is geworden aan vakinhoud of doordat leerkrachten gerichter vragen zijn gaan stellen aan leerlingen). Met behulp van een *mixed methods* onderzoek zou er meer zicht kunnen komen of deze aspecten effect hebben gehad op de kwaliteit van instructie. Mocht dit het geval zijn, dan kunnen deze aspecten in het vervolg van de nascholing nog meer worden benadrukt. Zo kan het programma naast effectiever ook efficiënter gemaakt worden. Andere suggesties voor vervolgonderzoek met betrekking tot het efficiënter dan wel effectiever maken van het programma zijn het uitvoeren van nascholing door de schoolleider of IB'er (zo sluit het programma beter aan bij de deelnemende leerkrachten en middenbouwteams) en specifieke aandacht voor subgroepen van leerlingen. Bij dit laatste moet worden gedacht aan zowel leerlingen die moeite hebben met begrijpend lezen als ook sterke lezers die vermoedelijk onvoldoende worden uitgedaagd.

Praktische aanbevelingen die in het kader van dit proefschrift worden gegeven hebben betrekking op het opleiden en nascholen van leerkrachten. De combinatie van aandacht voor vakspecifieke instructie en opbrengstgericht werken wordt als veelbelovend geacht, zo blijkt ook uit (onder andere) de review van Yoon et al. (2007) naar effectieve nascholing. Voor toekomstige nascholingen wordt een combinatie van deze componenten daarom aanbevolen. Daarnaast wordt het aangeraden om de inhoud van het nascholingsprogramma te integreren in de lerarenopleiding om zo de nodige basiskennis en –vaardigheden op het gebied van instructie voor begrijpend lezen, opbrengstgericht werken en het werken met doelen te stimuleren bij aankomende leerkrachten.

5. Slotwoord

De resultaten van deze dissertatie zijn waardevol voor de onderwijswetenschap omdat hiermee wordt aangetoond dat bevindingen uit onderzoek gebruikt kunnen worden ten behoeve van verbetering in de dagelijkse praktijk (zie ook Borko, 2004). In de woorden van Shulman over de meerwaarde van toegepast onderwijsonderzoek: “[these studies] evoke images of the possible (...) not only documenting that it can be done, but also laying out at least one detailed example of how it was organized, developed, and pursued” (Shulman, 1983, p. 495).

References

Aarnoutse, C. (1991). Begrijpend lezen in het basisonderwijs [Reading comprehension in primary school]. In P. Reitsma, & M. Walraven (Eds.), *Instructie in begrijpend lezen* (pp. 129-142). Delft, the Netherlands: Eburon.

Aarnoutse, C. (1992). Tekstgericht onderwijs in begrijpend lezen [Text-oriented teaching in reading comprehension]. In L. Verhoeven (Ed.), *Handboek lees- en schrijfdidactiek. functionele geletterdheid in basis- en voortgezet onderwijs* (pp. 151-166). Amsterdam, the Netherlands: Swets & Zeitlinger.

Aarnoutse, C., & Weterings, A. (1995). Onderwijs in begrijpend lezen [Education in reading comprehension]. *Pedagogische Studiën, 72*, 82-101.

Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Clarifying differences between reading skills and reading strategies. *The Reading Teacher, 61*(5), 364-373.

Andreassen, R., & Braten, I. (2011). Implementation and effects of explicit reading comprehension instruction in fifth-grade classrooms. *Learning and Instruction, 21*(4), 520-537.

Armbruster, B. B., Lehr, F., & Osborn, J. (2010). *Put reading first: The research building blocks for teaching children to read, kindergarten through grade 3*. Jessup, MD: National Institute for Literacy.

Baumann, J. F. (1988). Direct instruction reconsidered. *Journal of Reading, 31*(8), 712-718.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*(1), 137-172.

Bishop, A. R., Berryman, M. A., Wearmouth, J. B., & Peter, M. (2012). Developing an effective education reform model for indigenous and other minoritized students. *School Effectiveness and School Improvement, 23*(1), 49-70.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-63.

Black, P., & Wiliam, D. (1998b). Inside the black box. *Phi Delta Kappan, 80*(2), 139-148.

Bodovski, K., & Youn, M. J. (2011). The long term effects of early acquired skills and behaviors on young children's achievement in literacy and mathematics. *Journal of Early Childhood Research, 9*(1), 4-19.

Bond, G., Dykstra, R., Clymer, T., & Summers, E. G. (1997). The cooperative research program in first-grade reading instruction. *Reading Research Quarterly, 32*(4), 348-427.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8), 3-15.

Borko, H., Borko, H., & Koellner, K. (2010). Contemporary approaches to teacher professional development. In E. Baker, B. McGaw & P. Peterson (Eds.), *International encyclopedia of education* (3rd ed., pp. 548-555). Oxford, England: Elsevier Scientific Publications.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*(2), 125-230.

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of success for all. *American Educational Research Journal, 44*(3), 701-731.

Buijs, K., Klep, J., Noteboom, A., & Klein Tank, M. (2008). *TULE - rekenen / wiskunde: Inhouden en activiteiten bij de kerndoelen van 2006* [TULE - mathematics: contents and activities for the core objectives of 2006]. Enschede, the Netherlands: SLO.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the national teacher examinations. *Journal of Educational Measurement, 27*(2), 145-163.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31-72.

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis, 33*(3), 378-398.

Cito. (2003). *Categorieënoverzicht rekenen-wiskunde.* [Overview of mathematical sub-domains]. Arnhem, the Netherlands: Cito.

Cizek, G. J., & Bunch, M. B. (2006). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE Publications.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education, 12*(4), 343-366.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, D. K., & Ball, D. L. (1990). Policy and practice: An overview. *Educational Evaluation and Policy Analysis, 12*(3), 233-239.

Collins Block, C., & Lacina, J. (2009). Comprehension instruction in kindergarten through grade three. In S. E. Israel, & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 494). New York, NY: Routledge.

Council of Education. (2007). *Versteviging van kennis in het onderwijs II* [Reinforcement of knowledge in education II]. Utrecht, the Netherlands: Onderwijsraad.

Cozby, P. C. (2003). *Methods in behavioral research*. Boston, NY: McGraw-Hill.

Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving data: How high-performing school systems use data to improve instruction for elementary students.* Los Angeles, CA: University of Southern California.

de Boer, H., Bosker, R. J., & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology, 102*(1), 168-179.

de Jager, B. (2002). Teaching reading comprehension: A comparison of direct instruction and cognitive apprenticeship on comprehension skills and metacognition (Doctoral dissertation, University of Groningen, the Netherlands). Retrieved from http://dissertations.ub.rug.nl/FILES/faculties/gmw/ 2002 /b.de.jager/ thesis.pdf

de Jager, B., Reezigt, G. J., & Creemers, B. P. M. (2002). The effects of teacher training on new instructional behaviour in reading comprehension. *Teaching and Teacher Education, 18*(7), 831-842.

Deci, E. L. (2009). Large-scale school reform as viewed from the self-determination theory perspective. *Theory and Research in Education, 7*(2), 244-252.

Deming, W. E. (1986). *Out of the crisis: Quality, productivity and competitive position*. Cambridge, MA: Cambridge University Press.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181-199.

Deunk, M. I., van Kuijk, M. F., & Bosker, R. J. (in press). The effect of small group discussion on cutoff scores during standard setting. *Applied Measurement in Education.*

Doolaard, S., & Harms, T. (2013). *Omgaan met excellente leerlingen in de dagelijkse onderwijspraktijk* [Dealing with gifted students in daily educational practice]. Groningen, the Netherlands: University of Groningen.

Droop, M., van Elsäcker, W., & Voeten, R. (2012). *Effecten van de BLIKSEM-aanpak in groep 5 & 6 van het basisonderwijs* [Effects of the BLIKSEM-approach in third and fourth grade in primary school]. Nijmegen, the Netherlands: Expertisecentrum Nederlands.

Expert group Continuous Learning Progression. (2008). *Over de drempels met taal en rekenen* [Crossing the thresholds of language and mathematics]. Enschede, the Netherlands: Expertgroep Doorlopende Leerlijnen Taal en Rekenen.

Expertise Centre for the Dutch Language. (2010). *Van leerlijnen naar tussendoelen: Begrijpend lezen.* [From learning paths to short-term goals: reading comprehension]. Nijmegen, the Netherlands: Expertisecentrum Nederlands.

Feenstra, H., Kleintjes, F., Kamphuis, F., & Krom, R. (2010). *Wetenschappelijke verantwoording begrijpend lezen groep 3 t/m 6* [Scientific account for the reading comprehension tests grade 1 to 4]. Arnhem, the Netherlands: Cito.

Feenstra, H., Krom, R., & van Berkel, S. (2007a). *Begrijpend lezen groep 4: Handleiding* [Reading comprehension grade 2: manual]. Arnhem, the Netherlands: Cito.

Feenstra, H., Krom, R., & van Berkel, S. (2007b). *Begrijpend lezen groep 5: Handleiding* [Reading comprehension grade 3: manual]. Arnhem, the Netherlands: Cito.

Filipiak, P. (2006). De didactiek van het hardopdenkend (voor) lezen [the didactics of modeling]. *JSW, 90*, 20-23.

Fisher, D., Frey, N., & Lapp, D. (2008). Shared readings: Modeling comprehension, vocabulary, text structures, and text features for older readers. *The Reading Teacher, 61*(7), 548-556.

Freudenthal Institute, SLO & KPC. (2010). *Rekenlijn* [Mathematics line]. Retrieved from http://www.fi.uu.nl/rekenlijn/

Fuchs, L. S., Fuchs, D., & Deno, S. L. (1985). Importance of goal ambitiousness and goal mastery to student achievement. *Exceptional Children, 52*(1), 63-71.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Effects of alternative goal structures within curriculum-based measurement. *Exceptional Children, 55*(5), 429-438.

Fuchs, L. S., Fuchs, D., Stuebing, K., Fletcher, J. M., Hamlett, C. L., & Lambert, W. (2008). Problem solving and computational skill: Are they shared or distinct aspects of mathematical cognition? *Journal of Educational Psychology, 100*(1), 30-47.

Fuhrman, S. F. (2001). *From the Capitol to the classroom: Standards-based reform in the States*. Chicago, IL: the National Society for the Study of Education.

Fullan, M. (2001). *The new meaning of educational change*. New York, NY: Teachers College Press.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., & Jones, W. (2008). *The impact of two professional development interventions on early reading instruction and achievement*. Washington, DC: Institute of Education Sciences.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective: Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915-945.

Gay, G., & Kirkland, K. (2003). Developing cultural critical consciousness and self-reflection in preservice teacher education. *Theory into Practice, 42*(3), 181-187.

Good, T. L., & Brophy, J. E. (2003). *Looking in classrooms*. Boston, NY: Allyn & Baker.

Griffin, C. C., & Jittendra, A. K. (2009). Word problem-solving instruction in inclusive third-grade mathematics classrooms. *The Journal of Educational Research, 102*(3), 187-202.

Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching: Theory and Practice, 8*(3), 381-391.

Guthrie, J. T., Wigfield, A., Barbosa, P., Perencevich, K. C., Taboada, A., Davis, M. H., et al. (2004). Increasing reading comprehension and engagement through concept-oriented reading instruction. *Journal of Educational Psychology, 96*(3), 403-423.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (pp. 433-470). Washington, DC: American Council on Education.

Harris, M. J., & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin, 97*(3), 363-386.

Harskamp, E. (2010). *Programmeringsstudie rekenonderzoek in het primair onderwijs.* [Program research on studies in mathematics in primary school]. Den Haag, the Netherlands: NWO.

Hattie, J. A. C., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Heinrich, C., Maffioli, A., & Vazquez, G. (2010). *A primer for applying propensity score matching* (Technical Notes No. IDB-TN-161). Retrieved from Office of Strategic Planning and Development Effectiveness, Inter-American Development Bank website: http://www.iadb.org/en/publications/publication-detail,7101.html?id=8240%20

Herman, J. L., Osmundson, E., & Silver, D. (2010). *Capturing quality in formative assessment practice: Measurement challenges* (No. 770). Los Angeles, CA: CRESST.

Hill, H. C. (2007). Learning in the teaching workforce. *The Future of Children, 17*(1), 111-127.

Houtveen, A. A. M. (2002). *Begrijpend leesonderwijs dat werkt. Evaluatie van het adaptieve schoolverbeteringsproject "Kwaliteitsverbetering Begrijpend Lezen"* [Reading comprehension instruction that works. Evaluation of the adaptive school improvement project "Improvement of Reading Comprehension Quality"]. Utrecht, the Netherlands: Onderwijskunde/ISOR Onderwijsresearch.

Houtveen, A. A. M., & Mijs, D. (2004). *Evaluatie van Bossche stedelijk project FLEXIT: Een vierde tussenstand* [Evaluation of the urban city project FLEXIT: The fourth report]. Utrecht, the Netherlands: Onderwijskunde/ISOR Onderwijsresearch.

Houtveen, A. A. M., Mijs, D., Vernooy, K., & Roelofs, E. (2000). *Omgaan met verschillen bij beginnend lezen* [Dealing with differences in beginning readers]. Delft, the Netherlands: Eburon.

Houtveen, A. A. M., & Van de Grift, W. (2007). Reading instruction for struggling learners. *Journal of Education for Students Placed at Risk, 12*(4), 405-424.

Houtveen, A. A. M., & Van de Grift, W. (2012). Improving reading achievements of struggling learners. *School Effectiveness and School Improvement, 23*(1), 71-93.

Houtveen, A. A. M., Van de Grift, W., & Creemers, B. P. M. (2004). Effective school improvement in mathematics. *School Effectiveness and School Improvement, 15*(3-4), 337-376.

Huffman, D., & Kalnin, J. (2003). Collaborative inquiry to make data-based decisions in schools. *Teaching and Teacher Education, 19*(6), 569-580.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement, 63*(4), 584-601.

Inspectorate of Education. (2007). *De staat van het onderwijs: Jaarverslag 2005-2006* [The state of education: Year report 2005-2006]. Utrecht, the Netherlands: Onderwijsinspectie.

Inspectorate of Education. (2008). *De staat van het onderwijs. jaarverslag 2006-2007* [The state of education. Year report 2006-2007]. Utrecht, the Netherlands: Onderwijsinspectie.

Inspectorate of Education. (2010a). *Opbrengstgericht werken in het basisonderwijs* [Data-driven teaching in elementary schools]. Utrecht, the Netherlands: Onderwijsinspectie.

Inspectorate of Education. (2010b). *De staat van het onderwijs: Onderwijsverslag 2008-2009* [The state of education: Year report 2008-2009]. Utrecht, the Netherlands: Onderwijsinspectie.

Inspectorate of education. (2011). *Automatiseren bij rekenen-wiskunde. Een onderzoek naar het automatiseren van basisbewerkingen rekenen-wiskunde in het basisonderwijs.* [Automation in mathematics. A study on automation of arithmetic skills in primary school]. Utrecht, the Netherlands: Onderwijsinspectie.

Inspectorate of Education. (2011). *De staat van het onderwijs: Onderwijsverslag 2009-2010* [The state of education: Year report 2009-2010]. Utrecht, the Netherlands: Onderwijsinspectie.

Inspectorate of Education. (2012). *De staat van het onderwijs: Onderwijsverslag 2010-2011* [The state of education: Year report 2010-2011]. Utrecht, the Netherlands: Onderwijsinspectie.

Inspectorate of Education. (2013). *De staat van het onderwijs: Onderwijsverslag 2011-2012.* [The state of education: Year report 2011-2012.]. Utrecht, the Netherlands: Onderwijsinspectie.

Jacobse, A. E., & Harskamp, E. (2011). *A meta-analysis of the effects of instructional interventions on student's mathematics achievement.* Groningen, the Netherlands: GION.

Janssen, J., van der Schoot, F., & Hemker, B. (2005). *Balans van het reken-wiskundeonderwijs aan het einde van de basisschool* [Assessing performance in mathematics at the end of primary school]. Arnhem, the Netherlands: Cito/PPON.

Janssen, J., Verhulst, N., Engelen, R., & Scheltens, R. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS rekenen-wiskunde voor groep 3 t/m groep 8* [Scientific account for the mathematics tests grade 1 to 6]. Arnhem, the Netherlands: Cito.

Jimerson, S. R., & Kaufman, A. M. (2003). Reading, writing, and retention: A primer on grade retention research. *The Reading Teacher, 56*(7), 622-635.

Jussim, L., Eccles, J., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (28th ed., pp. 281-388). San Diego, CA: Academic Press.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*(2), 131-155.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mawwah, NJ: Lawrence Erlbaum Associates.

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Designing and evaluating standard-setting procedures for licensure and certification tests. *Advances in Health Sciences Education, 4*(3), 195-207.

Karantonis, A., & Sireci, S. G. (2006). The Bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4-12.

Kelcey, B. (2011). Assessing the effects of teachers' reading knowledge on students' achievement using multilevel propensity score stratification. *Educational Evaluation and Policy Analysis, 33*(4), 458-482.

Kirsch, I. S. (2002). *Reading for change: Performance and engagement across countries: Results from PISA 2000*. Paris, France: OECD.

Klieme, E., & Maag Merki, K. (2008). Introduction of educational standards in German-speaking countries. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 305-314). Cambridge, MA [etc.]: Hogrefe.

Kooiman, M. C., Hofman, R. H., Doolaard, S., & Guldemond, H. (2005). *Adaptief onderwijs in scholen voor speciaal basisonderwijs* [Adaptive special needs education]. Groningen, the Netherlands: GION.

Kordes, J., Bolsinova, M., Limpens, G., & Stolwijk, R. (2013). *PISA resultaten 2012: Praktische kennis en vaardigheden van 15-jarigen. Nederlandse uitkomsten van het Programme for International Student Assessment (PISA) op het gebied van leesvaardigheid, wiskunde en natuurwetenschappen in het jaar 2012* [PISA results 2012: skills and knowledge of 15-years olds. Results of Dutch students in PISA in regard to reading, mathematics and science in the year 2012]. Arnhem, the Netherlands: Cito.

Lai, M. K., & McNaughton, S. (2013). Analysis and discussion of classroom and achievement data to raise student achievement. In K. Schildkamp, & M. K. Lai (Eds.), *Data-based decision making in education: Challenges and opportunities* (pp. 23-47). Dordrecht, the Netherlands: Springer.

Lai, M. K., & Schildkamp, K. (2013). Data-based decision making: An overview. In K. Schildkamp, M. K. Lai & L. Earl (Eds.), *Data-based decision making in education: Challenges and opportunities* (pp. 9-21). Dordrecht, the Netherlands: Springer.

Lauer, P., Snow, D., Martin-Glenn, M., Van Buhler, R., Stoutemyer, K., & Snow-Renner, R. (2005). *The influence of standards on K-12 teaching and student learning: A research synthesis*. Aurora, CO: McRel.

Ledoux, G., Blok, H., & Boogaard, M. (2009). *Opbrengstgericht werken: Over de waarde van meetgestuurd werken* [Data use: The value of data-driven teaching]. Amsterdam, the Netherlands: SCO Kohnstamm Instituut.

Leenders, Y. G., Naafs, F. G., Oord, I. J., & Veenman, S. (2010). *Effectieve instructie: Leren lesgeven met het activerende directe instructiemodel* [Effective instruction: Learning to teach using the activiting Direct Instruction model]. Amersfoort, the Netherlands: CPS.

Liang, L. A., & Dole, J. A. (2006). Help with teaching reading comprehension: Comprehension instructional frameworks. *The Reading Teacher, 59*(8), 742-753.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Lloyd, D. N. (1978). Prediction of school failure from third-grade data. *Educational and Psychological Measurement, 38*(4), 1193-1200.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance.* Englewood Cliffs, NJ: Prentice-Hall.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*(9), 705-717.

Loman, E., & Marreveld, M. (2010). Plezier in begrijpend lezen: Themanummer. [Joy in comprehensive reading: Special issue] *Didactief, 8*, 1-16.

Madon, S., Jussim, L., & Eccles, J. (1997). In search of the powerful self-fulfilling prophecy. *Journal of Personality and Social Psychology, 72*(4), 791-809.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332-1361.

McGinty, D. (2005). Illuminating the" black box" of standard setting: An exploratory qualitative study. *Applied Measurement in Education, 18*(3), 269-287.

McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology, 46*(3), 235-261.

McMillan, J. H., & Schumacher, S. (1989). *Research in education, a conceptual introduction* (2nd ed.). Glennview, Il: Scott, Foresman and Company.

Meelissen, M. R. M., & Drent, M. (2008). *TIMMS-2007 Nederland. Trends in leerprestaties in exacte vakken in het basisonderwijs.* [TIMMS-2007 the Netherlands. Trends in performance in the exact sciences in primary school]. Enschede, the Netherlands: Universiteit Twente.

Meelissen, M. R. M., Netten, A., Drent, M., Punter, R. A., Droop, M., & Verhoeven, L. (2012). *PIRLS-en TIMSS-2011: Trends in leerprestaties in lezen, rekenen en natuuronderwijs* [PIRLS and TIMSS 2011: Trends in student performance in reading, mathematics and science]. Nijmegen, the Netherlands: Radboud Universiteit Nijmegen.

Meijer, J., & Ledoux, G. (2011). *Gebruikersvriendelijke leerlingvolgsystemen in het primair onderwijs* [User-friendliness of student monitoring systems in primary education]. Amsterdam, the Netherlands: SCO Kohnstamm Instituut.

Ministry of Education. (2008). *Beleidsreactie doorlopende leerlijnen taal en rekenen* [Policy response continuous learning trajectories language arts and mathematics]. Den Haag, the Netherlands: Ministerie van OCW.

Ministry of Education. (2009). *Toewijzing en invoering referentieniveaus taal en rekenen: Een evenwichtige ambitie* [Allocation and implementation attainment targets language arts and maths]. Den Haag, the Netherlands: Ministerie van OCW.

Ministry of Education. (2010). *Invoering referentieniveaus taal en rekenen* [Implementation performance standards language arts and mathematics]. Den Haag, the Netherlands: Ministerie van OCW.

Ministry of Education. (2011). *Voortgangsrapportage implementatie referentiekader taal en rekenen* [Progress report on the implementation of performance standards for language arts and mathematics]. Den Haag, the Netherlands: Ministerie van OCW.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahway, NJ: Lawrence Erlbaum Associates.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters: The junior years*. Wells, England: Open Books.

Muijs, D., & Reynolds, D. (2011). *Effective teaching: Evidence and practice*. London, England: SAGE.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development.

Norcini, J. J., & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education, 10*(1), 39-59.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research, 78*(1), 33-84.

OECD. (1995). *Performance standards in education: In search of quality*. Paris, France: OECD Publishing.

OECD. (2005). *Teachers matter: Attracting, developing and retaining effective teachers*. Paris, France: OECD Publishing.

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19*(3), 228-242.

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation, 35*(2-3), 95-101.

Peterson, C. H., Schulz, E. M., & Engelhard, G. (2011). Reliability and validity of Bookmark-based methods for standard setting: Comparisons to Angoff-based methods in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 30*(2), 3-14.

Plake, B. S. (2008). Standard setters: Stand up and take a stand! *Educational Measurement: Issues and Practice, 27*(1), 3-9.

Pressley, M. (1998). *Comprehension strategies instruction. literacy for all: Issues in teaching and learning*. New York, NY: Guilford Press.

Pressley, M., & Harris, K. R. (1990). What we really know about strategy instruction. *Educational Leadership, 48*(1), 31-34.

Primary Education Council (Producer). (2009). *Kijken naar opbrengsten* [Looking at results] [DVD]. Utrecht, the Netherlands.

Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2011). *MLwiN version 2.23*. Bristol, England: Centre for Multilevel Modeling.

Rasbash, J., Steele, F., Browne, W., & Goldstein, H. (2012). *A user's guide to MLwiN: Version 2.26*. Bristol, England: Centre for Multilevel Modeling.

Ravitch, D. (1995). *National standards in American schools: A citizen's guide*. Washington, DC: The Brookings Institution.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119-157). Thousand Oaks, CA: SAGE Publications.

Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., & Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Educational Research Journal, 48*(2), 462-501.

Roeber, E. D. (1999). Standards initiatives and American educational reform. In G. J. Cizek (Ed.), *Handbook of educational policy* (pp. 151-181). San Diego, CA: Academic Press.

Roeleveld, J., & Béguin, A. (2009). *Normering van referentieniveaus in het basisonderwijs* [norming performance standards in primary education]. Amsterdam, the Netherlands: SCO-Kohnstamm Instituut.

Rosenbaum, P. R. (2009). *Design of observational studies*. New York, NY: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38.

Rosenthal, R. (1987). " Pygmalion" effects: Existence, magnitude, and social importance. *Educational Researcher, 16*(9), 37-41.

Rosenthal, R., & Fode, K. L. (1963). The effect of experimenter bias on the performance of the albino rat. *Behavioral Science, 8*(3), 183-189.

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review, 3*(1), 16-20.

Royal Netherlands Academy of Arts and Sciences. (2009). *Rekenonderwijs op de basisschool. Analyse en sleutels tot verbetering* [mathematics education in primary school. Analyses and keys for improvement]. Amsterdam, the Netherlands: KNAW.

Rubie Davies, C. M. (2006). Teacher expectations and student self-perceptions: Exploring relationships. *Psychology in the Schools, 43*(5), 537-552.

Rubie Davies, C. M., Hattie, J., & Hamilton, R. (2006). Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology, 76*(3), 429-444.

Ruijsenaars, A. J. J. M., Van Luit, J. E. H., & Van Lieshout, E. C. D. M. (2004). *Rekenproblemen en dyscalculie* [Math problems and dyscalculia]. Rotterdam, the Netherlands: Lemniscaat.

Sammons, P., Hillman, J., & Mortimore, P. (1997). Key characteristics of effective schools: A review of school effectiveness research. In J. White, & M. Barber (Eds.), *Perspectives on school effectiveness and school improvement* (pp. 77-124). London, England: Institute of Education.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford, England: Pergamon.

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education, 26*(3), 482-496.

Schildkamp, K., Lai, M. K., & Earl, L. (2013). *Data-based decision making in education: Challenges and opportunities*. Dordrecht, the Netherlands: Springer.

Schnellert, L. M., Butler, D. L., & Higginson, S. K. (2008). Co-constructors of data, co-constructors of meaning: Teacher professional development in an age of accountability. *Teaching and Teacher Education, 24*(3), 725-750.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition and sense-making in mathematics. In D. Grouws (Ed.), *Handbook for research on mathematics teaching and learning* (pp. 334-370). New York, NY: MacMillan.

Seifert, T. (2004). Understanding student motivation. *Educational Research, 46*(2), 137-149.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, NY: Houghton Mifflin Company.

Sheskin, D. (2004). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: Chapman & Hall/CRC.

Shulman, L. S. (1983). Autonomy and obligation: The remote control of teaching. In L. S. Shulman, & G. Sykes (Eds.), *Handbook of teaching and policy* (pp. 484-504). New York, NY: Longman.

Siegel, S. (1988). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill Book Company.

Skaalvik, E. M., & Skaalvik, S. (2007). Dimensions of teacher self-efficacy and relations with strain factors, perceived collective teacher efficacy, and teacher burnout. *Journal of Educational Psychology, 99*(3), 611-625.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15-21.

Slavin, R. E. (2008). Perspectives on evidence-based research in education - What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5-14.

Slavin, R. E., Cheung, A., Holmes, G. C., Schools, A. C. P., Madden, N. A., & Chamberlain, A. (2013). Effects of a data-driven district reform model on state assessments. *American Educational Research Journal, 50*(2), 371-396.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research, 78*(3), 427-515.

Slavin, R. E., Madden, N. A., Chambers, B., & Haxby, B. (2009). *Two million children: Success for all*. Thousand Oaks, CA: Corwin Press.

Slavin, R. E., Madden, N. A., Dolan, L. J., Wasik, B. A., Ross, S., Smith, L., et al. (1996). Success for all: A summary of research. *Journal of Education for Students Placed at Risk, 1*(1), 41-76.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, England: SAGE Publications.

Snow, C. E. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: Rand Corporation.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academies Press.

Stattin, H., & Klackenberg-Larsson, I. (1993). Early language and intelligence development and their relationship to future criminal behavior. *Journal of Abnormal Psychology, 102*(3), 369-378.

Stoeldraijer, J., & Forrer, M. (2012). *Effectiever en efficiënter werken aan begrijpend lezen* [Working on reading comprehension skills more effectively and efficiently]. Utrecht, the Netherlands: PO-raad, School aan Zet.

Stoeldraijer, J., & Vernooy, K. (2007). Geen begrijpend lezen in groep 4?! Analyse van 5 methoden [No teaching of reading comprehension in grade 2?! Analysis of 5 curricular text books]. *Basisschoolmanagement, 21*(2), 10-16.

Sutherland, S. (2004). Creating a culture of data use for continuous improvement: A case study of an Edison project school. *American Journal of Evaluation, 25*(3), 277-293.

Swanson, H. L., & Hoskyn, M. (1998). Experimental intervention research on students with learning disabilities: A meta-analysis of treatment outcomes. *Review of Educational Research, 68*(3), 277-321.

Sweet, A. P., & Snow, C. E. (2003). *Rethinking reading comprehension*. New York, NY: Guilford Press.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston, NY: Allyn and Bacon.

Taylor, N. (2009). Standards-based accountability in South Africa. *School Effectiveness and School Improvement, 20*(3), 341-356.

Thoemmes, F. (2012). *Propensity score matching in SPSS*. Tübingen, Germany: University of Tübingen, Center for Educational Science and Psychology.

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., et al. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted, 27*(2/3), 119-145.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

van Berkel, S., Krom, R., Heesters, K., Van der Schoot, F., & Hemker, B. (2007). *Balans van het lees-onderwijs halverwege de basisschool* [Assessing performance in reading halfway primary school]. Arnhem, the Netherlands: Cito/PPON.

van de Mortel, K., & Förrer, M. (2010). *Lezen.. denken.. begrijpen!* [Reading.. Thinking.. Understanding!]. Amersfoort, the Netherlands: CPS.

van den Craats, J. (2007). *Rekenvaardigheden op de basisschool. Discussiestuk ten dienste van de Expertgroep Doorlopende Leerlijnen.* [Mathematical skills at primary school. Discussion paper for the Expert Group Continuous Learning Progression]. Retrieved from University of Amsterdam, the Netherlands website: http://staff.science.uva.nl/~craats/RekenenBasisschool.pdf

van de Grift, W., van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling van de pedagogisch-didactische vaardigheid van leraren in het basisonderwijs [Development in teaching skills]. *Pedagogische Studiën, 88*(6), 416-432.

van der Hoeven-van Doornum, Voeten, M. J. M., & Jungbluth, P. (1989). The effect of aspiration levels set by the teachers for their pupils on learning achievement. In B. P. M. Creemers, T. Peters & D. Reynolds (Eds.), *School effectiveness and school improvement* (pp. 231-239). Amsterdam, the Netherlands: Swets & Zeitlinger.

van der Kleij, F. M., & Eggen, T. J. H. M. (2013). Interpretation of the score reports from the computer program LOVS by teachers, internal support teachers and principals. *Studies in Educational Evaluation, 39*(3), 144-152.

van der Schoot, F. (2009). Cito variation on the bookmark method. In Council of Europe (Ed.), *Manual for relating language examinations to the common European framework of reference for languages* (pp. 1-16). Strasbourg, France: COE.

van Elsäcker, W. (2002). Development of reading comprehension: The engagement perspective. (Doctoral dissertation, University of Nijmegen, the Netherlands). Retrieved from http://dare.ubn.kun.nl/bitstream/2066/19128/1/19128_deveofrec.pdf

van Groenenstijn, M., Borghouts, C., & Janssen, C. (2011). *Protocol ernstige reken-wiskundeproblemen en dyscalculie.* [Protocol for severe mathematics problems and dyscalculia]. Assen, the Netherlands: Van Gorcum.

van Keer, H., & Verhaeghe, J. P. (2005). Comparing two teacher development programs for innovating reading comprehension instruction with regard to teachers' experiences and student outcomes. *Teaching and Teacher Education, 21*(5), 543-562.

Veenman, S., Leenders, Y., Meyer, P., & Sanders, M. (1993). Leren lesgeven met het directe instructiemodel [Learning to teach with the direct instruction model]. *Pedagogische Studiën, 70*, 2-16.

Verhoeven, L. (1991). Begrijpend lezen in de aanvangsfase: Rol van coherentie, inferentie en anafora [Reading comprehension in the early years: The role of coherence, inference and anafora]. In P. Reitsma, & M. Walraven (Eds.), *Instructie in begrijpend lezen* (pp. 113-127). Delft, the Netherlands: Eburon.

Vernooy, K. (2005). *Elke leerling een competente lezer!* [Each pupil a competent reader!]. Amersfoort, the Netherlands: CPS.

Wayman, J. C., Midgley, S., & Stringfield, S. (2006). Leadership for data-based decision making: Collaborative educator teams. In A. Danzig, K. Borman, B. Jones & B. Wright (Eds.), *Learner centered leadership. policy, research, and practice.* (pp. 189-205). Mahwah, NJ: Lawrence Erlbaum Associates.

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher, 37*(8), 469-479.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Washington, DC: Council of Chief State School Officers.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Skapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Young, V. (2006). Teachers' use of data: Loose coupling, agenda setting, and team norms. *American Journal of Education, 112*(4), 521-548.

About the author

Mechteld van Kuijk received her Bachelor's degree in Educational Sciences in 2006 and her Master's degree in Educational Sciences (cum laude) in 2007 at the University of Groningen. She continued studying at the University of Groningen, completing the Research Master 'Human Behavior in Social Contexts' in 2009, while combining her studies with a job as a school advisor and junior researcher at the school advisory centre 'Timpaan Onderwijs'. After completing her studies, Mechteld started her PhD-project at the Groningen Institute for Educational Research (GION). During her PhD, she was the Junior Researcher (JURE) coordinator for the EARLI's Special Interest Group (SIG) on Educational Effectiveness from 2009 to 2012 and she was the chair of the committee organizing the Educational Effectiveness SIG's JURE Pre-conference which was hosted in Zürich, Switzerland, in August of 2012. She also worked as a visiting researcher at the Centre for Evaluation and Assessment, University of Pretoria, South-Africa, in September and October of 2012. Mechteld is currently employed at the GION as a researcher and teacher. She is working on the Success for All - Nederland project, adapting this effective educational program to the Dutch context.

Dankwoord (acknowledgements in Dutch)

Dit proefschrift had ik nooit tot stand kunnen brengen zonder verschillende personen die ik graag hiervoor bedank. Allereerst wil ik mijn promotor Roel Bosker en dagelijks begeleider Marjolein Deunk bedanken voor hun ondersteuning de afgelopen jaren. Roel, dank voor de waardevolle input over de invulling van het onderzoek, het analyseren van de data en het schrijven van het proefschrift. Ik heb ontzettend veel van je geleerd. Bovendien combineer je het werk van een gedreven wetenschapper met ruimte voor *small talk* en grappen, iets wat ik zeer in je bewonder en wat me geregeld heeft doen lachen (totdat ik richting de einddatum van het project "het toch liever over mijn proefschrift wilde hebben"). Marjolein, jouw inzet bij de invulling en bij de uitvoering van het project zijn van enorme meerwaarde geweest. Daarbij heb je niets anders dan vertrouwen naar mij uitgestraald (je capoeira-naam "Confiança" is treffend gekozen door je medespelers). Dank voor je kritische blik en empatisch vermogen. Naast mijn begeleidingsteam is de bijdrage van Simone Doolaard erg waardevol geweest binnen het onderzoeksproject. Simone, dank voor het delen van jouw ervaring en het meedenken in alle aspecten van het onderzoek doen in scholen. In het kader van het onderzoeksproject wil ik ook graag de Cito-collega's Charlotte Jacobs en Jasmijn Oude Oosterik bedanken voor de plezierige samenwerking. En boven al ben ik de scholen die hebben deelgenomen aan het Streef-project zeer erkentelijk. Het werken met schoolleiders, intern begeleiders en leerkrachten vond ik ontzettend inspirerend, en het heeft mijn passie voor het onderzoek doen in en het werken met basisscholen alleen maar doen toenemen.

De afdeling Onderwijskunde/GION aan de RUG is er één waar ik met veel plezier werk en waar meerdere collega's een bijdrage hebben geleverd in de totstandkoming van dit proefschrift: dank voor de hulp met het koppelen van datasets (Henk), het meehelpen met de dataverzameling (studentassistenten Henriette, Els, Eva en Marloes) en de gesprekken over wetenschappelijk relevante en minder relevante zaken (het "lunch-rondje"). In het bijzonder wil ik het secretariaat bedanken. Vera, Stephanie en Sonja, jullie hulp (van het helpen met het verzenden van de Streef-post tot het nakijken van toetsen) was en is ontzettend waardevol en erg prettig. Daarnaast bedank ik graag *vrollega's* Annemieke, Mayra, Coby, Lieneke en Anouk voor de afleiding en de steun. Vooral paranimfen Lieneke (mijn Streef-maatje) en Anouk (mijn TIER-maatje) wil ik hier in de spotlight zetten. Wat ben ik blij dat jullie op deze dag, bij de verdediging van het proefschrift en tegelijkertijd het eind van het promotietraject, naast mij staan. De gesprekken op onze kamer, in de auto onderweg naar scholen, op reis naar conferenties; ik had nooit verwacht dat ik ooit zulke bewonderenswaardige en begripvolle kamergenoten zou hebben of dat ik zo vaak zo hard zou moeten lachen tijdens mijn werk. Jullie zijn absoluut van goud.

I would also like to thank colleagues from other universities whom I have met along the PhD-road (ICO, JURE, ORD, and the Educational Effectiveness SIG) and who have helped me in the

process of becoming a researcher. Furthermore, I would like to say ‘thank you’ to the researchers at the CEA (University of Pretoria, South Africa). My stay abroad at your institute has been a valuable research and life experience: it was truly eye-opening and heart-warming. In addition, I would like to acknowledge the importance of my Research Master friends. Mayra, Elisa, Annika, Lobke, and Marii, you are an inspiration to me both inside and outside the academic arena. Met name Mayra en Elisa, wat was het fijn om de laatste PhD-loodjes samen met jullie af te kunnen leggen. Onze reis in Chili was een fantastische afsluiting van deze periode.

Tot slot: vrienden, (schoon)familie en mijn lief Jasper - dank voor het vertrouwen. Mijn ouders en broers wil ik graag bedanken: jullie hebben mij altijd gestimuleerd om mijn best te (blijven) doen, en dat is mede de reden dat ik hier nu sta. En Jasper, jouw humor, nuchterheid en mooie liedjes maken me blij en geven mij steun; niet alleen tijdens dit promotie-proces maar al vanaf het moment dat ik je ken. Ik kijk er naar uit om deze gebeurtenis, het verdedigen van mijn proefschrift, met jullie (mijn dierbaren) te mogen delen en vieren. Ik zeg: feest!

ICO Dissertation series

In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the ICO Partner Universities: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, University of Amsterdam, University of Twente, Utrecht University, VU University Amsterdam, and Wageningen University, and formerly University of Groningen (until 2006), Radboud University Nijmegen (until 2004), and Tilburg University (until 2002). The University of Groningen, University of Antwerp, University of Ghent, and the Erasmus University Rotterdam have been 'ICO 'Network partner' in 2010 and 2011. From 2012 onwards, these ICO Network partners are full ICO partners, and from that period their dissertations will be added to this dissertation series.

List update: January, 2013 (the list will be updated every year in January)

234. Elffers, L. (14-12-2011). *The transition to post-secondary vocational education: Students' entrance, experiences, and attainment*. Amsterdam: University of Amsterdam.

235. Van Stiphout, I.M. (14-12-2011). *The development of algebraic proficiency.* Eindhoven: Eindhoven University of Technology.

236. Gervedink Nijhuis, C.J. (03-2-2012) *Culturally Sesitive Curriculum Development in International Cooperation.* Enschede: University of Twente.

237. Thoonen, E.E.J. (14-02-2012) *Improving Classroom Practices: The impact of Leadership School Organizational Conditions, and Teacher Factors.* Amsterdam: University of Amsterdam

238. Truijen, K.J.P (21-03-2012) *Teaming Teachers. Exploring factors that influence effective team functioning in a vocational education context.* Enschede: University of Twente.

239. Maulana, R.M. (26-03-2012) *Teacher-student relationships during the first year of secondary education. Exploring of change and link with motivation outcomes in The Netherlands and Indonesia.* Groningen: University of Groningen.

240. Lomos, C. (29-03-2012) *Professional community and student achievement.* Groningen: University of Groningen.

241. Mulder, Y.G. (19-04-2012) *Learning science by creating models.* Enschede: University of Twente.

242. Van Zundert, M.J. (04-05-2012) *Optimising the effectiveness and reliability of reciprocal peer assessment in secondary education.* Maastricht: Maastricht University.

243. Ketelaar, E. (24-05-2012) *Teachers and innovations: on the role of ownership, sense-making, and agency*. Eindhoven: Eindhoven University of Technology.

244. Logtenberg, A. (30-5-2012) *Questioning the past. Student questioning and historical reasoning.* Amsterdam: University of Amsterdam.

245. Jacobse, A.E. (11-06-2012) *Can we improve children's thinking?* Groningen: University of Groningen.

246. Leppink, J. (20-06-2012) *Propositional manipulation for conceptual understanding of statistics.* Maastricht: Maastricht University.

247. Van Andel, J (22-06-2012) *Demand-driven Education. An Educational-sociological Investigation.* Amsterdam: VU University Amsterdam.

248. Spanjers, I.A.E. (05-07-2012) *Segmentation of Animations: Explaining the Effects on the Learning Process and Learning Outcomes.* Maastricht: Maastricht University.

249. Vrijnsen-de Corte, M.C.W. *Researching the Teacher-Researcher. Practice-based research in Dutch Professional Development Schools.* Eindhoven: Eindhoven University of Technology.

250. Van de Pol, J.E. (28-09-2012) *Scaffolding in teacher-student interaction. Exploring, measuring promoting and evaluating scaffolding.* Amsterdam: University of Amsterdam.

251. Phielix, C. (28-09-2012) *Enhancing Collaboration through Assessment & Reflection.* Utrecht: Utrecht University.

252. Peltenburg, M.C. (24-10-2012) *Mathematical potential of special education students.* Utrecht: Utrecht University.

253. Doppenberg, J.J. (24-10-2012) *Collaborative teacher learning: settings, foci and powerful moments.* Eindhoven: Eindhoven University of Technology.

254. Kenbeek, W.K. (31-10-2012) *Back to the drawing board. Creating drawing or text summaries in support of System Dynamics modeling.* Enschede: University of Twente.

255. De Feijter, J.M. (09-11-2012) *Learning from error to improve patient safety.* Maastricht: Maastricht University.

256. Timmermans, A.C. (27-11-2012) *Value added in educational accountability: Possible, fair and useful?* Groningen: University of Groningen.

257. Van der Linden, P.W.J. (20-12-2012) *A design-based approach to introducing student teachers in conducting and using research.* Eindhoven: Eindhoven University of Technology.